D.V. BEKLEMISHEV

COURS DE GÉOMÉTRIE ANALYTIQUE ET D'ALGÈBRE LINÉAIRE

D. BEKLÉMICHEV

COURS DE GÉOMÉTRIE ANALYTIQUE ET D'ALGÈBRE LINÉAIRE



ÉDITIONS MIR · MOSCOU

TABLE DES MATIÈRES

PRÉFACE À L'ÉDITION FRANÇAISE

Le livre a son origine dans les cours professés par l'auteur pendant plusieurs années à l'Institut physico-technique de Moscou. Les chapitres I à X traitent du programme obligatoire de première année. On commence donc par les notions premières de la géométrie analytique dans le plan et dans l'espace, puis on passe à l'algèbre linéaire. L'exposé de la géométrie analytique vectorielle, qui en soi est utile aux futurs ingénieurs, constitue une bonne introduction à l'algèbre linéaire.

Puisque le cours est destiné aux futurs ingénieurs physiciens, il n'y a aucune raison de présenter les résultats sous une forme trop générale. C'est ainsi, par exemple, que pour le corps de base on considère le corps des réels ou le corps des complexes ; en définissant les tenseurs, on n'introduit pas le produit tensoriel des espaces. Vu l'importance des applications, on ne se limite pas aux aspects géométriques et l'on recourt constamment aux matrices. Les principaux instruments sont les transformations élémentaires des matrices.

Le volume du cours fondamental interdit de s'étendre sur les applications de l'algèbre linéaire aux problèmes scientifiques et techniques. Les cinq derniers chapitres du livre représentent un cours facultatif consacré à l'application correcte des méthodes de l'algèbre linéaire. Ils sont conçus de manière à laisser à l'étudiant une entière liberté dans le choix de la matière : les chapitres XIII, XIV et XV sont pratiquement indépendants l'un de l'autre, bien que certains passages des chapitres XIV et XV, traitant des méthodes de calculs, impliquent d'étudier le chapitre XIII. Le théorème de décomposition spectrale d'une fonction de matrice qui fait l'objet du chapitre XII sert souvent de référence dans les chapitres suivants. Le théorème de Jordan n'est pas utilisé.

Il existe de nombreux traités de théorie de programmation linéaire qui n'exigent pratiquement aucune formation mathématique. Dans ce livre, l'auteur a voulu présenter la programmation linéaire de manière à satisfaire le lecteur possédant un certain langage en mathématiques.

Dans les chapitres XI à XV, l'exposé est parfois informatif, nombre de résultats sont donnés sans démonstration. Dans ces cas, on renvoie aux sources contenant la démonstration.

L'auteur s'est efforcé, sans sacrifier à la clarté ou à la rigueur, de rendre le livre aussi concis que possible. Il a de même essayé dans la mesure du possible de faire partager au lecteur son admiration pour l'élégance de l'algèbre linéaire. Au lecteur de juger s'il y a réussi.

CHAPITRE PREMIER

ALGÈBRE VECTORIELLE

§ 1. Vecteurs

1. Remarques préliminaires. Les premiers chapitres de ce livre peuvent être considérés comme le prolongement du cours de géométrie enseigné à l'école secondaire. On sait que chaque discipline mathématique s'appuie sur un système de propositions qui n'exigent pas de démonstration et qu'on appelle *axiomes.* Le lecteur peut trouver la liste complète d'axiomes de géométrie ainsi que les raisonnements détaillés sur le rôle des axiomes en mathématiques dans le livre de N. Efimov [7] par exemple (les chiffres entre crochets sont des renvois à la liste d'ouvrages recommandés, insérée à la fin du livre).

On ne se pose pas pour but d'exposer les bases logiques du sujet, aussi s'appuie-t-on tout simplement sur les théorèmes démontrés dans le cours de géométrie élémentaire. Tous nos résultats ne peuvent donc être pris pour démontrés que dans la mesure où ces théorèmes le sont.

On ne tente pas non plus de définir les notions de base de la géométrie : point, droite, plan. Le lecteur curieux de leur introduction rigoureuse peut se référer au livre déjà mentionné de N. Efimov ; quant à nous, on admettra que ces notions, comme celles introduites dans le cours enseigné à l'école secondaire, sont connues du lecteur.

On suppose de même connues, dès l'école secondaire, la définition des nombres réels et leurs propriétés principales. (La théorie rigoureuse du nombre réel est exposée dans les manuels d'analyse mathématique.) On utilisera largement le fait que, l'unité de mesure une fois choisie, on est en droit d'associer à chaque segment un nombre réel positif appelé sa longueur. On admettra que *l'unité de mesure des longueurs est choisie une fois pour toutes* et, en parlant des longueurs de segments on ne mentionnera pas l'unité qui a servi à leur mesure.

2. Définition du vecteur. Un segment de droite est défini par deux points équivalents, ses extrémités. Mais on peut envisager un segment orienté défini par un couple ordonné de points. De ces points on sait lequel est le premier (origine) et lequel est le deuxième (extrémité).

DÉFINITION 1. On appellera *vecteur* un segment orienté (ou, ce qui revient au même, un couple ordonné de points). On rattachera également aux vecteurs le vecteur dit *nul* dont l'origine et l'extrémité coïncident.

Le sens sur le segment est habituellement marqué par une flèche. La notation littérale du vecteur est surmontée d'une flèche, par exemple, $\overrightarrow{AB}$ (la lettre correspondant à l'origine du vecteur est obligatoirement écrite la première). Dans les livres, les lettres désignant un vecteur sont souvent imprimées en caractères demi-gras, par exemple **a**. Le vecteur nul est noté **0** ou simplement 0.

La distance entre l'origine et l'extrémité du vecteur est appelée sa *longueur* (ou bien *module* ou *valeur absolue*). La longueur du vecteur est notée $|\mathbf{a}|$ ou $|\overrightarrow{AB}|$.

Les vecteurs sont dits *colinéaires* s'ils sont portés par une même droite ou des droites parallèles, bref, s'il existe une droite à laquelle ils sont parallèles. Les vecteurs sont dits *coplanaires* s'il existe un plan auquel ils sont parallèles.

Le vecteur nul est considéré colinéaire à tout vecteur, car il n'a pas de direction déterminée. Sa longueur est évidemment nulle.

DÉFINITION 2. Deux vecteurs sont dits *égaux* s'ils sont colinéaires, de même sens et ont des longueurs égales.

Il découle directement de cette définition qu'en choisissant un point arbitraire A' on est en mesure de construire un (et un seul) vecteur $\overrightarrow{A'B'}$ égal à un vecteur donné $\overrightarrow{AB}$, ou comme on dit, de transporter le vecteur $\overrightarrow{AB}$ au point A' (fig. 1).

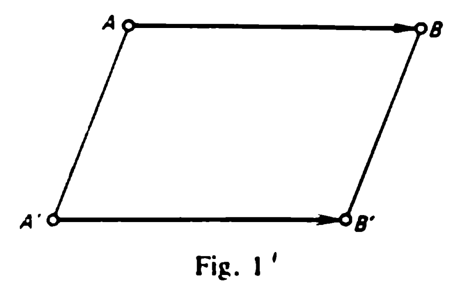


Fig. 1

3. Autre définition du vecteur. Remarquons que la notion d'égalité des vecteurs diffère sensiblement de celle des nombres. Tout nombre n'est égal qu'à lui-même, autrement dit, deux nombres égaux sont considérés comme un seul et même nombre en toute circonstance. Quant aux vecteurs, comme on le voit, il en est autrement : en vertu de la définition il existe des vecteurs différents mais égaux entre eux. Bien que dans la plupart des cas la nécessité de les distinguer ne se présente pas, il peut arriver (voir p. 23) qu'on attache plus d'importance, à un certain moment, au vecteur $\overrightarrow{AB}$ et non au vecteur $\overrightarrow{A'B'}$ qui lui est égal.

Pour simplifier la notion d'égalité des vecteurs (et lever certaines difficultés qui lui sont liées) on complique parfois la définition du vecteur. Tout en la formulant on n'utilisera pas cette définition compliquée. Afin d'éviter l'ambiguïté on écrira « Vecteur » (avec une lettre majuscule) pour désigner l'objet défini plus bas.

DÉFINITION 3. Soit un segment orienté. On appelle *Vecteur* l'ensemble de tous les segments orientés égaux au segment donné au sens de la définition 2.

Donc, chaque segment orienté définit un Vecteur. Il est aisé de constater que deux segments orientés définissent un même Vecteur si et seulement s'ils sont égaux. Pour les Vecteurs comme pour les nombres l'égalité signifie coïncidence : deux Vecteurs sont égaux si et seulement si c'est le même Vecteur.

Au cas d'une translation dans l'espace, tout point de l'espace et son image constituent un couple ordonné et définissent un segment orienté, tous ces segments orientés étant égaux au sens de la définition 2. Aussi la translation de l'espace peut-elle être identifiée avec le Vecteur qui réunit tous ces segments orientés.

On sait du cours élémentaire de physique que la force peut être représentée par un segment orienté, mais ne peut pas l'être par un Vecteur, car les forces figurées par des segments orientés égaux effectuent en général des actions différentes. (Si la force agit sur un corps élastique, le segment orienté qui la représente ne peut être transporté même le long de la droite qui le supporte.)

Ce n'est qu'une des raisons pour lesquelles à côté des Vecteurs, c'est-à-dire des ensembles (ou comme on dit des classes) de segments orientés égaux, on est obligé de tenir compte des représentants isolés de ces classes. Dans ces circonstances, l'utilisation de la définition 3 se complique de nombreuses restrictions. On se tiendra donc à la définition 1, et le lecteur comprendra chaque fois, d'après le sens général, s'il s'agit d'un vecteur bien défini ou d'un vecteur qui peut être remplacé par tout vecteur qui lui est égal.

En rapport avec la définition du vecteur il est utile d'élucider le sens de certains mots rencontrés dans les livres.

Au lieu de la définition 2 on peut introduire une autre définition de l'égalité des vecteurs selon laquelle les vecteurs sont égaux s'ils ont même longueur, même support et même sens. Dans ce cas le vecteur peut être transporté non pas en tout point de l'espace mais seulement le long de la droite qui le supporte. Avec ce concept d'égalité, les vecteurs sont dits *glissants* ou *localisés sur une droite*. En mécanique, la force agissant sur un solide parfait est représentée par un vecteur glissant.

On peut de même ne pas introduire de concept particulier d'égalité de vecteurs et considérer que chaque vecteur n'est égal qu'à lui-même et se caractérise, en plus de la longueur et de la direction dans l'espace, par le point d'application. Dans ce cas les vecteurs sont dits *liés* (ou *localisés en un point*). Comme il a été indiqué la force agissant sur un corps élastique est représentée par un vecteur lié.

S'il est nécessaire de souligner que l'égalité est comprise au sens de la définition 2, on dit que ce vecteur est *libre*. On représente par un vecteur libre, par exemple, la vitesse angulaire d'un corps.

4. Opérations linéaires sur des vecteurs. On appelle *opérations linéaires* l'addition des vecteurs et la multiplication du vecteur par un nombre. Rappelons leurs définitions.

DÉFINITION. Soient donnés deux vecteurs $\mathbf{a}$ et $\mathbf{b}$. Construisons les vecteurs $\overrightarrow{AB}$ et $\overrightarrow{BC}$ qui leur sont égaux (c'est-à-dire transportons l'extrémité de $\mathbf{a}$ et l'origine de $\mathbf{b}$ au même point B). Alors le vecteur $\overrightarrow{AC}$ est appelé *somme* des vecteurs $\mathbf{a}$ et $\mathbf{b}$ et est noté $\mathbf{a} + \mathbf{b}$.

Remarquons qu'en choisissant au lieu de B un autre point, par exemple, B' on obtiendrait comme somme un autre vecteur $\overrightarrow{A'C'}$ égal à $\overrightarrow{AC}$.

On appelle *addition* de vecteurs l'opération faisant correspondre à deux vecteurs leur somme.

DÉFINITION. On appelle *produit du vecteur* $\mathbf{a}$ *par un nombre réel* α tout vecteur $\mathbf{b}$ satisfaisant aux conditions suivantes :

a) $|\mathbf{b}| = |\alpha|\ |\mathbf{a}|$;

b) le vecteur **b** est colinéaire au vecteur **a** ;

c) les vecteurs **b** et **a** sont orientés dans le même sens si $\alpha > 0$, et dans les sens opposés si $\alpha < 0$. (Mais si $\alpha = 0$, il découle de la première condition que **b** = **0**.)

Le produit du vecteur **a** par le nombre α est noté $\alpha\mathbf{a}$.

La *multiplication d'un vecteur par un nombre* est une opération qui fait correspondre au vecteur et au nombre le produit du vecteur par ce nombre.

Dans le cours enseigné à l'école secondaire on a déduit les principales propriétés des opérations linéaires. Enumérons-les sans démonstration.

PROPOSITION 1. 1) *L'addition des vecteurs est commutative, c'est-à-dire que pour tous vecteurs* **a** *et* **b** *on a* $\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}$.

2) *L'addition des vecteurs est associative, c'est-à-dire que pour tous vecteurs* **a**, **b** *et* **c** *on a* $\mathbf{a} + (\mathbf{b} + \mathbf{c}) = (\mathbf{a} + \mathbf{b}) + \mathbf{c}$.

3) *L'addition du vecteur nul à tout vecteur* **a** *ne le modifie pas* : $\mathbf{a} + \mathbf{0} = \mathbf{a}$.

4) *Pour tout vecteur* **a** *le vecteur* $(-1)\mathbf{a}$ *est son opposé, c'est-à-dire* : $\mathbf{a} + (-1)\mathbf{a} = \mathbf{0}$.

5) *La multiplication d'un vecteur par un nombre est associative, c'est-à-dire que pour tous nombres* α *et* β *et tout vecteur* **a** *on a* $(\alpha\beta)\mathbf{a} = \alpha(\beta\mathbf{a})$.

6) *La multiplication d'un vecteur par un nombre est distributive par rapport à l'addition des nombres, c'est-à-dire que pour tous nombres* α *et* β *et tout vecteur* **a** *on a* $(\alpha + \beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$.

7) *La multiplication d'un vecteur par un nombre est distributive par rapport à l'addition des vecteurs, c'est-à-dire que pour tous vecteurs* **a** *et* **b** *et tout nombre* α *on a* $\alpha(\mathbf{a} + \mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$.

8) *La multiplication d'un vecteur par l'unité ne modifie pas ce vecteur* : $1\mathbf{a} = \mathbf{a}$.

Le vecteur opposé au vecteur **a** est noté $-\mathbf{a}$. On appelle *différence* des vecteurs **a** et **b** la somme du vecteur **a** et du vecteur opposé à **b**, c'est-à-dire le vecteur $\mathbf{a} + (-\mathbf{b})$ ou de façon condensée $\mathbf{a} - \mathbf{b}$.

La *soustraction* est une opération inverse de l'addition, qui fait correspondre à deux vecteurs leur différence : à partir de la somme de deux vecteurs $\mathbf{b} + \mathbf{x} = \mathbf{a}$ et de l'un des termes **b** on peut obtenir le deuxième terme $\mathbf{x} = \mathbf{a} - \mathbf{b}$ (fig. 2).

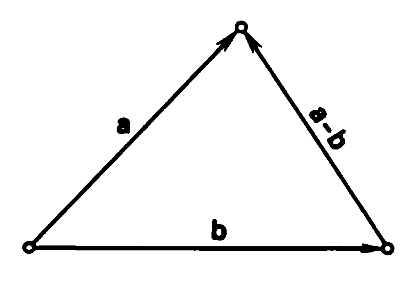


Fig. 2.

La soustraction est définie à l'aide de l'addition, aussi ne la considérera-t-on comme une opération indépendante. On ne distinguera pas non plus la division d'un vecteur par le nombre α, qu'on peut définir comme une multiplication par le nombre α^{-1}.

En se servant des opérations linéaires on est en mesure de composer des sommes de vecteurs multipliés par des nombres : $\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \ldots$ $\ldots + \alpha_k \mathbf{a}_k$. L'expression de ce type s'appelle *combinaison linéaire* de vecteurs, et les nombres qui y figurent ses *coefficients*.

Les propriétés d'opérations linéaires énoncées ci-dessus permettent de transformer les expressions constituées de combinaisons linéaires en se conformant aux règles habituelles d'algèbre : on peut chasser les parenthèses, réduire les termes semblables, transporter certains termes d'un membre de l'égalité dans l'autre en changeant leurs signes, etc.

Les huit propriétés d'opérations linéaires énoncées dans la proposition 1 constituent, en un certain sens, un ensemble complet de propriétés : elles suffisent pour effectuer tout calcul utilisant les opérations linéaires sur des vecteurs, sans avoir recours aux définitions de ces opérations. Ce fait présentera par la suite (ch. VI) une importance de principe.

Les combinaisons linéaires de vecteurs présentent les propriétés évidentes suivantes : si les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ sont colinéaires, toute combinaison linéaire de ces derniers leur est colinéaire ; si les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ sont coplanaires, toute combinaison linéaire de ces derniers leur est coplanaire. Cela découle aussitôt du fait que le vecteur $\alpha\mathbf{a}$ est colinéaire à $\mathbf{a}$ et que la somme $\mathbf{a} + \mathbf{b}$ est située dans le même plan que $\mathbf{a}$ et $\mathbf{b}$ et même sur la même droite si les vecteurs $\mathbf{a}$ et $\mathbf{b}$ sont colinéaires.

DÉFINITION. On appelle *base d'un espace* un triplet de vecteurs non coplanaires pris dans un certain ordre.

On appelle *base d'un plan* un couple de vecteurs non colinéaires de ce plan pris dans un certain ordre.

On appelle *base d'une droite* tout vecteur non nul de ce support.

Soulignons que les vecteurs de base d'un plan ne sont pas nuls, car si l'un d'eux l'était, ils seraient colinéaires. De même, deux vecteurs quelconques de base d'un espace ne sont jamais colinéaires, sinon tous les trois seraient coplanaires.

Si un vecteur est une combinaison linéaire de vecteurs, on dit qu'il est *décomposé* suivant ces vecteurs. Le plus souvent on envisage la décomposition d'un vecteur suivant les vecteurs de base.

DÉFINITION. Si $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ est une base de l'espace et $\mathbf{a} = \alpha_1 \mathbf{e}_1 +$ $+ \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3$, les nombres $\alpha_1, \alpha_2, \alpha_3$ sont appelés *composantes* (ou *coordonnées*) du vecteur $\mathbf{a}$ par rapport à la base considérée.

D'une façon analogue, sont définies les composantes du vecteur dans un plan et sur une droite. Les composantes du vecteur s'écrivent entre

parenthèses juste après la notation littérale du vecteur. Par exemple, $\mathbf{a}(1, 0, 1)$ signifie que les composantes du vecteur $\mathbf{a}$ par rapport à la base déterminée préalablement choisie sont 1, 0 et 1.

THÉORÈME 1. *Tout vecteur parallèle à une droite peut être décomposé suivant le vecteur de base de cette droite.*

Tout vecteur parallèle à un plan peut être décomposé suivant les vecteurs de base de ce plan.

Tout vecteur peut être décomposé suivant les vecteurs de base de l'espace.

Les composantes du vecteur se définissent dans chaque cas de façon univoque.

DÉMONSTRATION. La première assertion signifie que pour tout vecteur $\mathbf{a}$ colinéaire à un vecteur non nul $\mathbf{e}$ (base d'une droite) il existe un nombre α tel que $\mathbf{a} = \alpha\mathbf{e}$. Ce nombre est : soit $|\mathbf{a}| / |\mathbf{e}|$, soit $-|\mathbf{a}| / |\mathbf{e}|$ suivant que $\mathbf{a}$ et $\mathbf{e}$ sont de même sens ou de sens contraires.

La deuxième assertion signifie que pour tout vecteur $\mathbf{a}$ coplanaire à deux vecteurs non colinéaires $\mathbf{e}_1$ et $\mathbf{e}_2$ (base d'un plan) il existe des nombres α_1 et α_2 tels que $\mathbf{a} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2$. Pour fixer ces nombres, plaçons les origines des trois vecteurs au point O et menons par l'extrémité A du vecteur $\mathbf{a}$ la droite AP parallèle à $\mathbf{e}_2$ (fig. 3). On a alors $\overrightarrow{OA} = \overrightarrow{OP} + \overrightarrow{PA}$, $\overrightarrow{OP}$ étant colinéaire à $\mathbf{e}_1$ et $\overrightarrow{PA}$ à $\mathbf{e}_2$. (En particulier, l'un des vecteurs $\overrightarrow{OP}$ et $\overrightarrow{PA}$ peut s'avérer nul.) En vertu de la première assertion du théorème il existe un α_1 et un α_2 tels que $\overrightarrow{OP} = \alpha_1\mathbf{e}_1$ et $\overrightarrow{PA} = \alpha_2\mathbf{e}_2$. Il s'ensuit que $\overrightarrow{OA} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2$.

La troisième assertion du théorème signifie que pour tout vecteur $\mathbf{a}$ et des vecteurs non coplanaires $\mathbf{e}_1$, $\mathbf{e}_2$ et $\mathbf{e}_3$ il existe des nombres α_1, α_2 et α_3 pour lesquels $\mathbf{a} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \alpha_3\mathbf{e}_3$. Pour le démontrer, plaçons les origines de tous les vecteurs au même point O (fig. 4) et menons par l'extré-

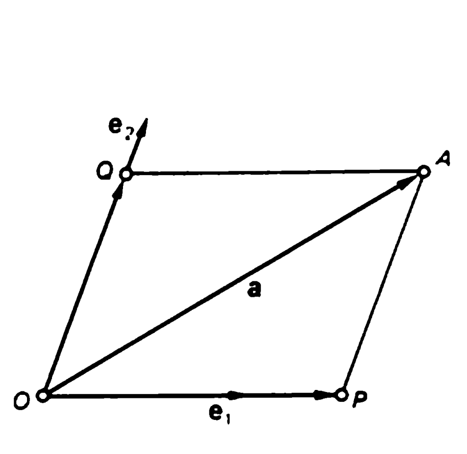


Fig. 3.

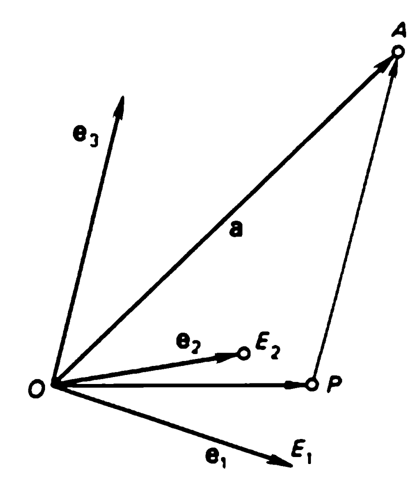


Fig. 4.

mité A du vecteur $\mathbf{a}$ la droite AP parallèle au vecteur $\mathbf{e}_3$. On a alors $\overrightarrow{OA} = \overrightarrow{OP} + \overrightarrow{PA}$, avec $\overrightarrow{PA}$ colinéaire à $\mathbf{e}_3$ et $\overrightarrow{OP}$ coplanaire à $\mathbf{e}_1$ et $\mathbf{e}_2$. En vertu des assertions déjà démontrées, on obtient les nombres α_1, α_2 et α_3 tels que $\overrightarrow{PA} = \alpha_3 \mathbf{e}_3$ et $\overrightarrow{OP} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2$. D'où découle directement la troisième assertion.

Supposons qu'un vecteur $\mathbf{a}$ est décomposé suivant les vecteurs de base de l'espace de deux manières : $\mathbf{a} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3$ et $\mathbf{a} = \beta_1 \mathbf{e}_1 + \beta_2 \mathbf{e}_2 + \beta_3 \mathbf{e}_3$. En retranchant la deuxième expression de la première, on obtient $(\alpha_1 - \beta_1)\mathbf{e}_1 + (\alpha_2 - \beta_2)\mathbf{e}_2 + (\alpha_3 - \beta_3)\mathbf{e}_3 = \mathbf{0}$. Si au moins une des différences entre parenthèses n'est pas nulle, on est en mesure de décomposer l'un des vecteurs de base suivant les autres. Par exemple, avec $\alpha_1 - \beta_1 \neq 0$, il vient

$$\mathbf{e}_1 = -\frac{\alpha_2 - \beta_2}{\alpha_1 - \beta_1}\mathbf{e}_2 - \frac{\alpha_3 - \beta_3}{\alpha_1 - \beta_1}\mathbf{e}_3,$$

ce qui est faux car les vecteurs de base ne sont pas coplanaires. La contradiction obtenue démontre l'unicité de la décomposition suivant les vecteurs de base de l'espace. De façon analogue on démontre l'unicité de la décomposition dans les autres cas. Le théorème est complètement démontré.

En reprenant la démonstration de la dernière partie du théorème, on remarquera qu'elle est également applicable à la proposition qui suit.

PROPOSITION 2. *Les vecteurs égaux possèdent les mêmes composantes.*

En géométrie analytique, les raisonnements géométriques sur les vecteurs se réduisent à des calculs où participent les composantes de ces vecteurs. Les deux propositions suivantes montrent comment on effectue les opérations connues sur les vecteurs au cas où sont données leurs composantes.

PROPOSITION 3. *Pour multiplier un vecteur par un nombre on multiplie tous ses composantes par ce nombre.*

En effet, si $\mathbf{a} = \alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3$, on a

$$\lambda \mathbf{a} = \lambda(\alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3) = (\lambda\alpha_1)\mathbf{e}_1 + (\lambda\alpha_2)\mathbf{e}_2 + (\lambda\alpha_3)\mathbf{e}_3.$$

PROPOSITION 4. *Pour additionner les vecteurs on procède à l'addition de leurs composantes respectives.*

En effet, si $\mathbf{a} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3$ et $\mathbf{b} = \beta_1 \mathbf{e}_1 + \beta_2 \mathbf{e}_2 + \beta_3 \mathbf{e}_3$, il vient

$$\mathbf{a} + \mathbf{b} = (\alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3) + (\beta_1 \mathbf{e}_1 + \beta_2 \mathbf{e}_2 + \beta_3 \mathbf{e}_3) =$$
$$= (\alpha_1 + \beta_1)\mathbf{e}_1 + (\alpha_2 + \beta_2)\mathbf{e}_2 + (\alpha_3 + \beta_3)\mathbf{e}_3.$$

5. Dépendance linéaire des vecteurs. La combinaison linéaire de plusieurs vecteurs est dite *triviale* si tous ses coefficients sont nuls. Certes, la

combinaison linéaire triviale de vecteurs est égale au vecteur nul. La combinaison linéaire est *non triviale* si au moins un de ses coefficients est différent de zéro.

DÉFINITION. Les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ sont dits *linéairement dépendants* s'il existe une combinaison linéaire non triviale de ces vecteurs qui est égale à zéro ; autrement dit s'il existe des coefficients $\alpha_1, \ldots, \alpha_k$ pour lesquels $\alpha_1 \mathbf{a}_1 + \ldots + \alpha_k \mathbf{a}_k = \mathbf{0}$ et $\alpha_1^2 + \ldots + \alpha_k^2 \neq 0$.

Dans le cas contraire, c'est-à-dire lorsque seule la combinaison linéaire triviale des vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ est nulle, ces vecteurs sont dits *linéairement indépendants*. Si les vecteurs sont linéairement indépendants, l'égalité $\alpha_1 \mathbf{a}_1 + \ldots + \alpha_k \mathbf{a}_k = \mathbf{0}$ entraîne $\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0$.

Signalons les propriétés suivantes de la dépendance linéaire.

Si l'un des vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ est nul, ces vecteurs sont linéairement dépendants. En effet, considérons une combinaison linéaire de ces vecteurs dans laquelle le coefficient du vecteur nul est 1, tandis que les autres coefficients sont égaux à 0. Cette combinaison linéaire n'est pas triviale et est égale à zéro.

Si l'on ajoute un ou plusieurs vecteurs $\mathbf{b}_1, \ldots, \mathbf{b}_j$ aux vecteurs linéairement dépendants $\mathbf{a}_1, \ldots, \mathbf{a}_k$, les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k, \mathbf{b}_1, \ldots, \mathbf{b}_j$ sont encore linéairement dépendants. En effet, à une combinaison non triviale nulle des vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ on peut adjoindre les vecteurs $\mathbf{b}_1, \ldots, \mathbf{b}_j$ munis des coefficients nuls.

PROPOSITION 5. *Les vecteurs sont linéairement dépendants si et seulement si l'un d'eux se décompose suivant les autres.*

DÉMONSTRATION. Supposons que $\mathbf{a}_1, \ldots, \mathbf{a}_k$ sont linéairement dépendants, c'est-à-dire qu'il existe des coefficients $\alpha_1, \ldots, \alpha_k$ tels que $\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \ldots + \alpha_k \mathbf{a}_k = \mathbf{0}$ et l'un au moins des coefficients, par exemple α_1, est différent de zéro. Dans ce cas, $\mathbf{a}_1$ est une combinaison linéaire des vecteurs $\mathbf{a}_2, \ldots, \mathbf{a}_k$. En effet, on peut écrire

$$\mathbf{a}_1 = -\frac{\alpha_2}{\alpha_1}\mathbf{a}_2 - \ldots - \frac{\alpha_k}{\alpha_1}\mathbf{a}_k.$$

Inversement, supposons qu'un des vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$, par exemple $\mathbf{a}_1$, se décompose suivant les autres, c'est-à-dire est une combinaison linéaire suivante :

$$\mathbf{a}_1 = \beta_2 \mathbf{a}_2 + \ldots + \beta_k \mathbf{a}_k.$$

Il s'ensuit immédiatement que la combinaison linéaire des vecteurs $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ à coefficients $-1, \beta_2, \ldots, \beta_k$ est égale au vecteur nul. Vu qu'elle n'est pas triviale, les vecteurs $\mathbf{a}_1, \ldots, \mathbf{a}_k$ sont linéairement dépendants. La proposition est démontrée.

La notion de dépendance linéaire jouera un grand rôle dans la suite. Pour l'instant on peut s'en dispenser vu la signification géométrique banale de la dépendance linéaire des vecteurs.

PROPOSITION 6. *Deux vecteurs colinéaires sont toujours linéairement dépendants, et inversement, deux vecteurs linéairement dépendants sont colinéaires.*

En effet, soient donnés deux vecteurs colinéaires. S'ils sont tous deux nuls, l'assertion est évidente ; si l'un des vecteurs n'est pas nul, l'autre se décompose suivant ce vecteur. Dans les deux cas les vecteurs sont linéairement dépendants.

Inversement, en vertu de la proposition 5, de deux vecteurs linéairement dépendants l'un s'exprime en fonction de l'autre et par suite, ils sont colinéaires.

PROPOSITION 7. *Trois vecteurs coplanaires sont toujours linéairement dépendants, et inversement, trois vecteurs linéairement dépendants sont coplanaires.*

DÉMONSTRATION. Supposons donnés trois vecteurs coplanaires. Considérons deux quelconques d'entre eux. S'ils sont colinéaires, donc linéairement dépendants, ils le seront encore avec le troisième vecteur. Par contre, si deux vecteurs ne sont pas colinéaires, le troisième se décompose suivant ces derniers, et les trois vecteurs sont linéairement dépendants en vertu de la proposition 5.

Inversement, de trois vecteurs linéairement dépendants l'un se décompose suivant les deux autres et, par suite, leur est coplanaire (et même colinéaire, s'ils sont colinéaires).

PROPOSITION 8. *Quatre vecteurs sont toujours linéairement dépendants.*

En effet, considérons trois quelconques de ces quatre vecteurs. S'ils sont coplanaires, donc linéairement dépendants, ils le sont encore avec le quatrième vecteur. S'ils ne sont pas coplanaires, le quatrième vecteur se décompose suivant les trois autres, d'où il découle que les quatre vecteurs sont linéairement dépendants.

§ 2. Systèmes de coordonnées

1. Système de coordonnées cartésiennes. Fixons dans l'espace un point O et considérons un point arbitraire M. On dit que le vecteur $\overrightarrow{OM}$ est le *rayon vecteur* du point M relativement au point O. Si outre le point O, on a choisi dans l'espace une base, il devient possible de faire correspondre au point M un triplet ordonné de nombres, *composantes* de son rayon vecteur.

Définition. On appelle *système de coordonnées cartésiennes* (ou *repère cartésien*) de l'espace l'ensemble constitué d'un point et d'une base.

Le point est dénommé *origine des coordonnées* ; les droites passant par l'origine des coordonnées et portant les vecteurs de base sont les *axes de coordonnées*. La première est l'*axe des abscisses*, la deuxième, l'*axe des ordonnées*, la troisième, l'*axe des cotes*. Les plans passant par les axes de coordonnées sont appelés *plans de coordonnées*.

Définition. Les composantes du rayon vecteur du point M relativement à l'origine des coordonnées sont appelées *coordonnées* du point M dans le système considéré de coordonnées.

La première coordonnée est l'*abscisse*, la deuxième, l'*ordonnée* et la troisième, la *cote*.

Les coordonnées cartésiennes dans le plan et sur la droite se définissent de façon analogue. Il va de soi que dans le plan un point ne possède que deux coordonnées (l'abscisse et l'ordonnée) et sur la droite qu'une seule.

Les coordonnées d'un point s'écrivent habituellement entre parenthèses après la lettre désignant le point. Par exemple, l'écriture A (2, 1/2) signifie que le point A du plan a pour coordonnées 2 et 1/2 par rapport au repère cartésien préalablement choisi dans ce plan (fig. 5).

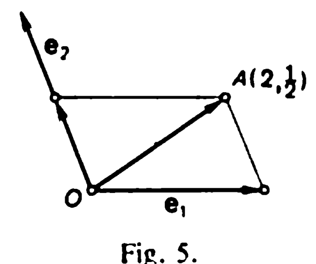


Fig. 5.

Il est aisé de voir qu'avec le repère donné les coordonnées d'un point sont définies de façon univoque. Réciproquement, étant donné un repère, à tout triplet ordonné de nombres correspond un point et un seul dont les coordonnées sont ces nombres. Le repère du plan définit une correspondance analogue entre les points du plan et les couples ordonnés de nombres.

Considérons deux points A et B dont les coordonnées par rapport au système de coordonnées cartésiennes $\{O, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ sont respectivement x_1, y_1, z_1 et x_2, y_2, z_2. Cherchons les composantes du vecteur $\overrightarrow{AB}$. Il est évident que $\overrightarrow{AB} = \overrightarrow{OB} - \overrightarrow{OA}$ (fig. 6). De par la définition des coordonnées les composantes des rayons vecteurs $\overrightarrow{OA}$ et $\overrightarrow{OB}$ sont (x_1, y_1, z_1) et (x_2, y_2, z_2). Selon la proposition 4 du § 1, le vecteur $\overrightarrow{AB}$ possède les composantes $(x_2 - x_1, y_2 - y_1, z_2 - z_1)$. On a ainsi démontré la proposition suivante.

Proposition 1. *Pour obtenir les composantes d'un vecteur il faut retrancher les coordonnées de son origine de celles de son extrémité.*

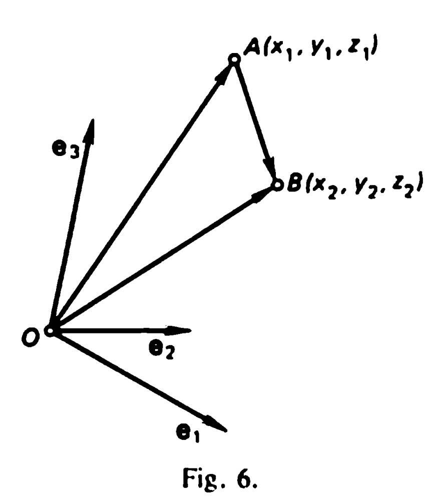


Fig. 6.

2. Partage d'un segment dans un rapport donné. Cherchons les coordonnées du point M situé sur le segment AB et le partageant dans le rapport λ/μ, c'est-à-dire satisfaisant à la condition

$$\frac{|AM|}{|MB|} = \frac{\lambda}{\mu}, \quad \lambda > 0, \quad \mu > 0$$

(fig. 7). Cette condition peut être écrite sous la forme

$$\mu\overrightarrow{AM} = \lambda\overrightarrow{MB}. \tag{1}$$

En désignant par (x_1, y_1, z_1) et (x_2, y_2, z_2) les coordonnées respectives des points A et B et par (x, y, z) celles du point M, décomposons les deux mem-

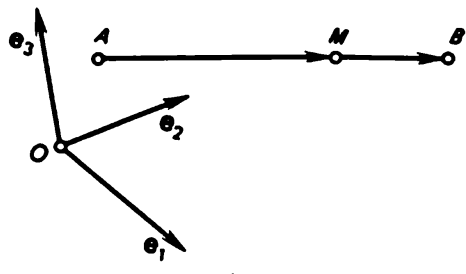


Fig. 7.

bres de l'égalité suivant les vecteurs de base et utilisons la proposition 1 pour obtenir les composantes des vecteurs $\overrightarrow{AM}$ et $\overrightarrow{MB}$. On a alors

$$\mu(x - x_1) = \lambda(x_2 - x),$$
$$\mu(y - y_1) = \lambda(y_2 - y),$$
$$\mu(z - z_1) = \lambda(z_2 - z).$$

Ces égalités permettent d'obtenir x, y et z vu que $\lambda + \mu \neq 0$. On obtient en définitive

$$x = \frac{\mu x_1 + \lambda x_2}{\lambda + \mu}, \quad y = \frac{\mu y_1 + \lambda y_2}{\lambda + \mu}, \quad z = \frac{\mu z_1 + \lambda z_2}{\lambda + \mu}. \tag{2}$$

Ces formules sont connues sous le nom de *formule de partage d'un segment dans un rapport donné.*

Si dans les formules (2) l'un des nombres λ ou μ est strictement négatif, il résulte de l'égalité (1) que le point $M(x, y, z)$ est situé sur la même droite mais à l'extérieur du segment AB, en le partageant dans le rapport $|\lambda/\mu|$. Aussi les formules (2) constituent-elles la solution d'un problème plus général ; elles permettent de trouver les coordonnées d'un point partageant le segment AB dans un rapport donné, intérieurement et extérieurement.

Dans le plan, le problème de partage d'un segment se résout de la même façon, à la seule différence que la base n'est constituée que de deux vecteurs et, par suite, des formules (2) il ne reste que deux.

3. Repère cartésien rectangulaire. De tous les repères cartésiens on utilise le plus souvent des repères qui forment une classe spéciale des repères cartésiens rectangulaires.

DÉFINITION. Une base est dite *orthonormée* si ses vecteurs sont unitaires et deux à deux orthogonaux. Le repère cartésien dont la base est orthonormée est appelé *repère cartésien rectangulaire.*

Il n'est pas difficile de vérifier que les coordonnées d'un point par rapport à un repère cartésien rectangulaire sont, en valeur absolue, égales aux distances de ce point jusqu'aux plans de coordonnées respectifs. Elles sont strictement positives ou négatives suivant que le point et l'extrémité du vecteur de base se trouvent d'un même côté ou de part et d'autre du plan de coordonnées auquel est perpendiculaire ce vecteur.

De façon analogue, on obtient les coordonnées d'un point par rapport à un repère cartésien rectangulaire dans le plan.

4. Système de coordonnées polaires. Les repères cartésiens ne sont pas les seuls repères qu'on utilise pour déterminer la position d'un point par rapport à une certaine image géométrique. Il existe beaucoup d'autres systèmes de coordonnées. On décrira ici quelques-uns d'entre eux.

On utilise souvent dans le plan le *système de coordonnées polaires.* Il est défini si est donné un point O appelé *pôle*, et une demi-droite l issue de O, appelée *axe polaire.* La position d'un point M est fixée par deux nombres : le *rayon* $r = |\overrightarrow{OM}|$ et l'angle φ entre l'axe polaire et le vecteur $\overrightarrow{OM}$. L'angle φ est appelé *angle polaire.* On le mesure en radians à partir de l'axe polaire dans le sens contraire à celui du mouvement des aiguilles d'une montre. Pour le pôle O on a $r = 0$ et φ est indéterminé. Pour les autres points, $r > 0$ et φ est déterminé au multiple de 2π près. Cela signifie que, par exemple, les couples de nombres (r, φ), $(r, \varphi + 2\pi)$ et, en général, $(r, \varphi + 2k\pi)$, où k est un entier quelconque, représentent les coordonnées polaires d'un même point (fig. 8).

On limite parfois la variation de l'angle polaire en introduisant diverses

conditions, par exemple : $0 \leqslant \varphi < 2\pi$ ou $-\pi < \varphi' \leqslant \pi$. Ce procédé supprime la non-univocité mais introduit d'autres inconvénients.

Soient donnés un système de coordonnées polaires et un couple ordonné (r, φ) de nombres dont le premier est positif. On peut faire correspondre à ce couple un point dont les coordonnées polaires sont ces nombres, à savoir : si $r = 0$, c'est le pôle qu'on fait correspondre, si $r > 0$, c'est le point dont le rayon vecteur est de longueur r et forme l'angle φ avec l'axe polaire. Ceci étant, aux couples (r, φ) et (r_1, φ_1) correspond un même point si $r = r_1$ et $\varphi - \varphi_1 = 2\pi k$, où k est un entier.

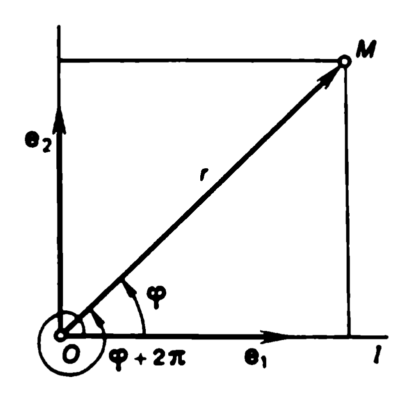


Fig. 8.

Choisissons dans le plan un repère cartésien rectangulaire en plaçant son origine au pôle O et en prenant pour vecteurs $\mathbf{e}_1$, $\mathbf{e}_2$ les vecteurs de longueur 1 dirigés respectivement le long de l et sous l'angle $\pi/2$ à l (l'angle étant mesuré dans le sens contraire au mouvement des aiguilles d'une montre). Il découle aisément de la figure 8 que les coordonnées cartésiennes du point s'expriment par ses coordonnées polaires de la manière suivante :

$$x = r\cos\varphi, \quad y = r\sin\varphi. \tag{3}$$

5. Coordonnées cylindriques et sphériques. Les systèmes de coordonnées cylindriques et sphériques sont une généralisation dans l'espace des systèmes de coordonnées polaires. Pour les uns comme pour les autres, la figure par rapport à laquelle on détermine la position d'un point est composée d'un point O, d'une demi-droite l issue de O et d'un vecteur $\mathbf{n}$ de longueur unité perpendiculaire à l. Par le point O on peut mener le plan P perpendiculaire au vecteur $\mathbf{n}$.

Soit un point M. Traçons par ce point la perpendiculaire MM' au plan P.

Les *coordonnées cylindriques* du point M sont représentées par trois nombres (r, φ, h). Les nombres r, φ sont les coordonnées polaires du point M' par rapport au pôle O et l'axe polaire l, tandis que h est la composante du vecteur $\overrightarrow{M'M}$ suivant $\mathbf{n}$. Elle est déterminée car ces vecteurs sont colinéaires (fig. 9).

Les *coordonnées sphériques* d'un point sont trois nombres (r, φ, θ). Elles se déterminent de la façon suivante : $r = |\overrightarrow{OM}|$, comme pour les coordonnées cylindriques φ est l'angle entre le vecteur $\overrightarrow{OM'}$ et la demi-droite l, tandis que θ est l'angle que fait le vecteur $\overrightarrow{OM}$ avec le plan P (fig. 10).

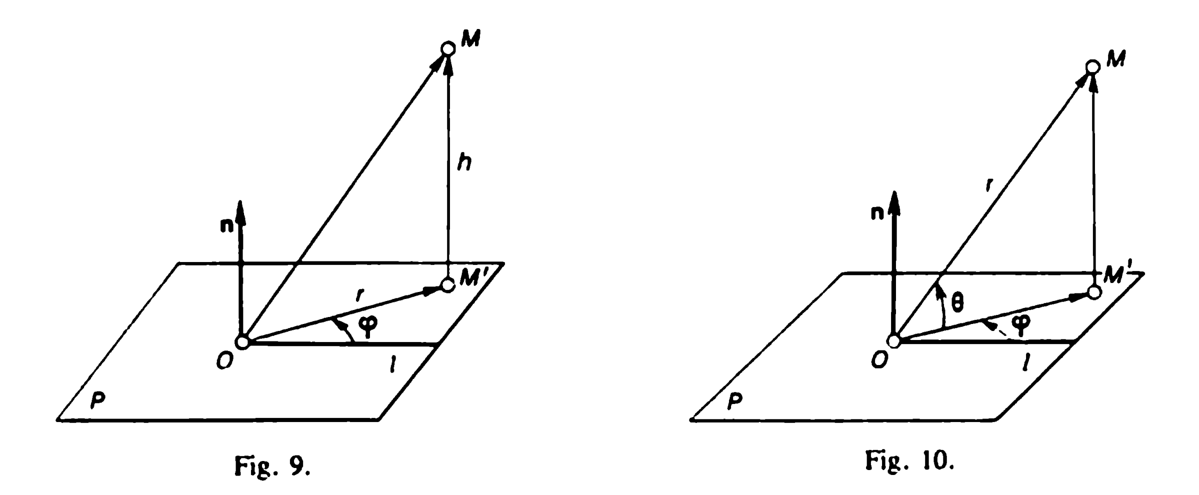


Fig. 9. Fig. 10.

§ 3. Produits scalaire et vectoriel

1. Produit scalaire. On entend par angle de deux vecteurs l'angle formé par deux vecteurs égaux aux vecteurs donnés et ayant une origine commune. Quelquefois, on indiquera le vecteur à partir duquel on mesure cet angle, et le sens dans lequel on le fait. En l'absence de telle indication, l'angle de deux vecteurs est celui qui est inférieur à π. Si l'angle est droit, les vecteurs sont dits *orthogonaux*.

DÉFINITION. On appelle *produit scalaire* de deux vecteurs le nombre égal au produit des longueurs de ces vecteurs par le cosinus de l'angle qu'ils forment. Si l'un des vecteurs est nul, l'angle n'est pas déterminé et le produit scalaire est égal à zéro par définition.

Le produit scalaire des vecteurs $\mathbf{a}$ et $\mathbf{b}$ est représenté par $(\mathbf{a}, \mathbf{b})$. On peut donc écrire

$$(\mathbf{a}, \mathbf{b}) = |\mathbf{a}|\ |\mathbf{b}| \cos \varphi,$$

où φ est l'angle des vecteurs $\mathbf{a}$ et $\mathbf{b}$. L'opération de multiplication scalaire jouit des propriétés évidentes suivantes :

1. *La multiplication scalaire est commutative, c'est-à-dire* $(\mathbf{a}, \mathbf{b}) = (\mathbf{b}, \mathbf{a})$ *pour tous vecteurs* $\mathbf{a}$ *et* $\mathbf{b}$.

2. $(\mathbf{a}, \mathbf{a}) = |\mathbf{a}|^2$ *pour tout vecteur* $\mathbf{a}$.

3. *Le produit scalaire est nul si et seulement si les facteurs sont orthogonaux ou si l'un d'eux au moins est nul.*

4. *Les vecteurs d'une base orthonormée satisfont aux relations*

$$(\mathbf{e}_1, \mathbf{e}_1) = (\mathbf{e}_2, \mathbf{e}_2) = (\mathbf{e}_3, \mathbf{e}_3) = 1,$$
$$(\mathbf{e}_1, \mathbf{e}_2) = (\mathbf{e}_2, \mathbf{e}_3) = (\mathbf{e}_3, \mathbf{e}_1) = 0.$$

PROPOSITION 1. *Si les vecteurs de base* $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ *sont orthogonaux, les*

composantes de tout vecteur **a** *s'obtiennent d'après les formules*

$$\alpha_1 = \frac{(\mathbf{a}, \mathbf{e}_1)}{|\mathbf{e}_1|^2}, \quad \alpha_2 = \frac{(\mathbf{a}, \mathbf{e}_2)}{|\mathbf{e}_2|^2}, \quad \alpha_3 = \frac{(\mathbf{a}, \mathbf{e}_3)}{|\mathbf{e}_3|^2}.$$

En particulier, si la base est orthonormée, on a

$$\alpha_1 = (\mathbf{a}, \mathbf{e}_1), \quad \alpha_2 = (\mathbf{a}, \mathbf{e}_2), \quad \alpha_3 = (\mathbf{a}, \mathbf{e}_3). \tag{1}$$

En effet, soit $\mathbf{a} = \mathbf{a}_1 + \mathbf{a}_2 + \mathbf{a}_3$, chaque terme étant colinéaire au vecteur de base respectif. On sait de la démonstration du théorème 1, § 1, que $\alpha_1 = \pm |\mathbf{a}_1| / |\mathbf{e}_1|$, le signe étant plus ou moins suivant que les vecteurs $\mathbf{a}_1$ et $\mathbf{e}_1$ sont de même sens ou de sens contraires. Or il découle de la figure 11 que $\pm |\mathbf{a}_1| = |\mathbf{a}| \cos \varphi_1$, où φ_1 est l'angle des vecteurs **a** et $\mathbf{e}_1$. Ainsi, $\alpha_1 = |\mathbf{a}| \cos \varphi_1 / |\mathbf{e}_1| = (\mathbf{a}, \mathbf{e}_1) / |\mathbf{e}_1|^2$. De façon analogue se calculent les autres composantes.

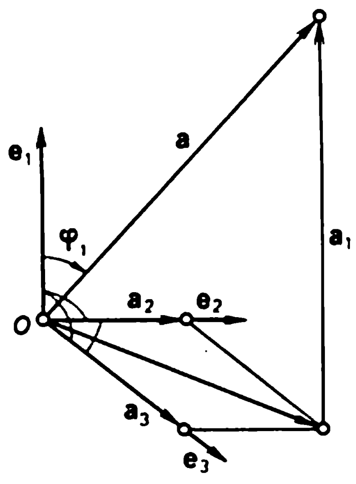


Fig. 11.

De la proposition 1 il découle que chacune des composantes du vecteur **a** par rapport à une base orthonormée est égale au produit de sa longueur par le cosinus de l'angle que ce vecteur fait avec le vecteur de base correspondant.

La propriété suivante est dénommée *linéarité du produit scalaire.*

PROPOSITION 2. *Quels que soient les vecteurs* **a**, **b**, **c** *et les nombres* α *et* β, *on a l'égalité* $(\alpha\mathbf{a} + \beta\mathbf{b}, \mathbf{c}) = \alpha(\mathbf{a}, \mathbf{c}) + \beta(\mathbf{b}, \mathbf{c})$. *En particulier,* $(\alpha\mathbf{a}, \mathbf{c}) = \alpha(\mathbf{a}, \mathbf{c})$ *et* $(\mathbf{a} + \mathbf{b}, \mathbf{c}) = (\mathbf{a}, \mathbf{c}) + (\mathbf{b}, \mathbf{c})$.

DÉMONSTRATION. Si $\mathbf{c} = \mathbf{0}$, l'assertion est évidente. Posons $\mathbf{c} \neq \mathbf{0}$. Convenons que **c** est le premier vecteur de base. Choisissons les deux autres de façon qu'ils soient orthogonaux entre eux et au premier vecteur. L'expression $(\alpha\mathbf{a} + \beta\mathbf{b}, \mathbf{c}) / |\mathbf{c}|^2$ est la première composante du vecteur $\alpha\mathbf{a} + \beta\mathbf{b}$. D'une façon analogue, $(\mathbf{a}, \mathbf{c}) / |\mathbf{c}|^2$ et $(\mathbf{b}, \mathbf{c}) / |\mathbf{c}|^2$ sont les premières composantes des vecteurs **a** et **b**. En vertu des propositions 4 et 3 du

§ 1, on a

$$(\alpha\mathbf{a} + \beta\mathbf{b}, \mathbf{c}) / |\mathbf{c}|^2 = \alpha(\mathbf{a}, \mathbf{c}) / |\mathbf{c}|^2 + \beta(\mathbf{b}, \mathbf{c}) / |\mathbf{c}|^2,$$

d'où l'égalité exigée.

On démontre aisément que la combinaison linéaire d'un nombre quelconque de vecteurs vérifie une formule analogue.

Profitant de la commutativité de la multiplication scalaire, on obtient de la proposition 2 l'identité suivante :

$$(\mathbf{a}, \beta\mathbf{b} + \gamma\mathbf{c}) = \beta(\mathbf{a}, \mathbf{b}) + \gamma(\mathbf{a}, \mathbf{c}).$$

Soit donnée une base orthonormée. Etudions comment le produit scalaire des vecteurs **a** et **b** s'exprime au moyen de leurs composantes (α_1, α_2, α_3) et (β_1, β_2, β_3). En s'appuyant sur la linéarité du produit scalaire par rapport au premier facteur, écrivons

$$(\mathbf{a}, \mathbf{b}) = (\alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \alpha_3\mathbf{e}_3, \mathbf{b}) = \alpha_1(\mathbf{e}_1, \mathbf{b}) + \alpha_2(\mathbf{e}_2, \mathbf{b}) + \alpha_3(\mathbf{e}_3, \mathbf{b}).$$

Mais en vertu de la proposition 1, les produits scalaires du vecteur **b** par les vecteurs de base sont égaux à ses composantes β_1, β_2 et β_3. On aboutit ainsi au théorème suivant.

THÉORÈME 1. *Si la base est orthonormée, le produit scalaire des vecteurs s'exprime au moyen de leurs composantes suivant la formule*

$$(\mathbf{a}, \mathbf{b}) = \alpha_1\beta_1 + \alpha_2\beta_2 + \alpha_3\beta_3. \tag{2}$$

Soulignons l'importance de l'exigence, dans ce théorème, d'une base orthonormée, car rapportée à toute autre base l'expression du produit scalaire en fonction des composantes est beaucoup plus compliquée. Aussi dans les problèmes où interviennent les produits scalaires emploie-t-on des bases orthonormées. Mais si pour des raisons quelconques il s'avère nécessaire de calculer le produit scalaire par rapport à une base non orthonormée, il faut multiplier scalairement les facteurs décomposés suivant les vecteurs de base et, en chassant les parenthèses, porter dans l'expression obtenue les produits scalaires connus des vecteurs de base.

Le théorème 1 permet d'exprimer la longueur du vecteur en fonction de ses composantes par rapport à une base orthonormée :

$$|\mathbf{a}| = \sqrt{\alpha_1^2 + \alpha_2^2 + \alpha_3^2}, \tag{3}$$

ainsi que l'angle de deux vecteurs en fonction de leurs composantes par rapport à une base orthonormée :

$$\cos\varphi = \frac{(\mathbf{a}, \mathbf{b})}{|\mathbf{a}|\,|\mathbf{b}|} = \frac{\alpha_1\beta_1 + \alpha_2\beta_2 + \alpha_3\beta_3}{\sqrt{(\alpha_1^2 + \alpha_2^2 + \alpha_3^2)(\beta_1^2 + \beta_2^2 + \beta_3^2)}}. \tag{4}$$

En se servant de la formule (3), on est en mesure de calculer la distance de deux points si sont données leurs coordonnées par rapport à un repère *cartésien rectangulaire*. En effet, supposons que les points A et B possèdent les coordonnées respectives (x, y, z) et (x_1, y_1, z_1). Alors la distance entre ces points est égale à

$$|\overrightarrow{AB}| = \sqrt{(x_1 - x)^2 + (y_1 - y)^2 + (z_1 - z)^2}. \qquad (5)$$

2. Orientation d'un triplet de vecteurs. Soient données deux bases orthonormées $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ et $\{\mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$. Peut-on faire coïncider ces bases à l'aide d'un déplacement? On peut évidemment transporter et imprimer un mouvement de rotation au vecteur $\mathbf{e}_1'$ de manière qu'il s'applique sur le vecteur $\mathbf{e}_1$. Dans ce cas, le plan des vecteurs $\mathbf{e}_2'$ et $\mathbf{e}_3'$, perpendiculaire à $\mathbf{e}_1'$, sera amené sur le plan des vecteurs $\mathbf{e}_2$ et $\mathbf{e}_3$, perpendiculaire à $\mathbf{e}_1$. Ensuite, par rotation, on peut faire coïncider les vecteurs $\mathbf{e}_2'$ et $\mathbf{e}_2$. Les vecteurs $\mathbf{e}_3'$ et $\mathbf{e}_3$ seront alors colinéaires. Ceci étant, ou bien ils se confondront, les bases coïncidant, ou bien ils seront opposées, les bases ne pouvant alors s'appliquer l'une sur l'autre.

Il découle de ce raisonnement que, si deux bases ne peuvent être amenées l'une sur l'autre, toute autre base s'applique soit sur la première, soit sur la deuxième. Ainsi, toutes les bases orthonormées se partagent en deux classes. Les bases d'une même classe se superposent, tandis que les bases des classes différentes ne se superposent pas. La base est dite *orientée à droite* (*directe*) ou *à gauche* (*rétrograde*) suivant que le vecteur $\mathbf{e}_1$ est orienté à droite ou à gauche, $\mathbf{e}_2$ s'éloignant de nous et le vecteur $\mathbf{e}_3$ se dirigeant vers le haut (fig. 12). L'une des classes comprend toutes les bases directes et l'autre toutes les bases rétrogrades. Cette définition s'étend à toutes les bases de la façon suivante.

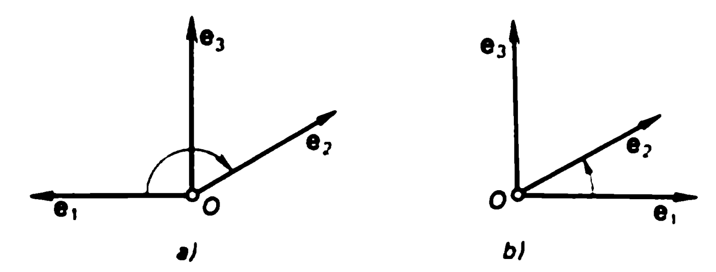


Fig. 12. a) Base rétrograde; b) base directe

DÉFINITION. Un triplet ordonné de vecteurs non coplanaires est dit *orienté à droite* (ou *direct*) si de l'extrémité du troisième vecteur la plus petite rotation du premier vecteur vers le deuxième s'observe dans le sens contraire au mouvement des aiguilles d'une montre. Dans le cas contraire, le triplet est dit *orienté à gauche* (ou *rétrograde*). (Les origines des vecteurs du triplet sont supposées confondues.)

3. Produit vectoriel. DÉFINITION. Soient donnés deux vecteurs **a** et **b**. Construisons un vecteur **c** satisfaisant aux conditions suivantes :

1) $|\mathbf{c}| = |\mathbf{a}|\,|\mathbf{b}| \sin \varphi$ *), où φ est l'angle entre **a** et **b** ;

2) le vecteur **c** est orthogonal aux vecteurs **a** et **b** ;

3) si **a** et **b** ne sont pas colinéaires, les vecteurs **a**, **b**, **c** forment un triplet direct de vecteurs.

Le vecteur ainsi construit est appelé *produit vectoriel* des vecteurs **a** et **b** et désigné par $[\mathbf{a}, \mathbf{b}]$. Les conditions mentionnées définissent le produit vectoriel à l'égalité près si les facteurs ne sont pas des vecteurs nuls. Si l'un au moins des facteurs est nul, le produit vectoriel est par définition le vecteur nul.

Il découle de la définition que le module du produit vectoriel de deux vecteurs non colinéaires est numériquement égal à l'aire du parallélogramme construit sur les facteurs (au cas où les facteurs ont une origine commune).

Le produit vectoriel est nul si et seulement si les facteurs sont colinéaires **).

EXEMPLE 1. Soit $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ une base orthonormée directe. Alors $[\mathbf{e}_1, \mathbf{e}_2] = \mathbf{e}_3$, $[\mathbf{e}_2, \mathbf{e}_3] = \mathbf{e}_1$, $[\mathbf{e}_3, \mathbf{e}_1] = \mathbf{e}_2$. Si $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$ est une base orthonormée rétrograde, $[\mathbf{f}_1, \mathbf{f}_2] = -\mathbf{f}_3$, $[\mathbf{f}_2, \mathbf{f}_3] = -\mathbf{f}_1$, $[\mathbf{f}_3, \mathbf{f}_1] = -\mathbf{f}_2$.

PROPOSITION 3. *La multiplication vectorielle est anticommutative, c'est-à-dire qu'on a toujours* $[\mathbf{a}, \mathbf{b}] = -[\mathbf{b}, \mathbf{a}]$.

En effet, il résulte de la définition que le module du produit vectoriel est indépendant de l'ordre des facteurs. De même, le vecteur $[\mathbf{a}, \mathbf{b}]$ est colinéaire au vecteur $[\mathbf{b}, \mathbf{a}]$. Toutefois, en permutant les facteurs, on est obligé de changer le sens du produit pour que soit remplie la condition 3) de la définition. En effet, si **a**, **b**, $[\mathbf{a}, \mathbf{b}]$ forment un triplet direct, **b**, **a**, $[\mathbf{a}, \mathbf{b}]$ forment alors un triplet rétrograde, et **b**, **a**, $-[\mathbf{a}, \mathbf{b}]$ de nouveau un triplet direct.

On reviendra plus bas au produit vectoriel mais, avant de le faire, on va introduire encore une opération.

4. Produit mixte. DÉFINITION. Le nombre $(\mathbf{a}, [\mathbf{b}, \mathbf{c}])$ est appelé *produit mixte des vecteurs* **a**, **b**, **c** et noté $(\mathbf{a}, \mathbf{b}, \mathbf{c})$.

PROPOSITION 4. *Le produit mixte des vecteurs non coplanaires* **a**, **b** *et* **c** *est égal en module au volume du parallélépipède construit sur les facteurs. Il est positif si le triplet* **a**, **b**, **c** *est direct et négatif s'il est rétrograde.*

---

*) $\sin \varphi \geqslant 0$, car $0 \leqslant \varphi \leqslant \pi$.

**) On a convenu de considérer que le vecteur nul est colinéaire à tout vecteur.

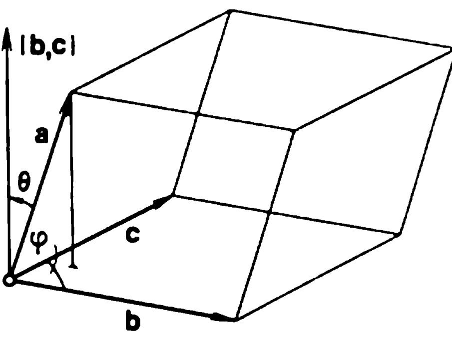


Fig. 13.

En effet, le volume du parallélépipède construit sur les vecteurs **a**, **b** et **c** est égal (fig. 13) au produit de l'aire de la base $|[\mathbf{b}, \mathbf{c}]|$ par la hauteur $|\mathbf{a}||\cos\theta|$, où θ est l'angle des vecteurs **a** et $[\mathbf{b}, \mathbf{c}]$. On peut donc écrire

$$V = |[\mathbf{b}, \mathbf{c}]|\ |\mathbf{a}|\ |\cos\theta| = |(\mathbf{a}, [\mathbf{b}, \mathbf{c}])| = |(\mathbf{a}, \mathbf{b}, \mathbf{c})|.$$

Ainsi, la première assertion est démontrée. Le signe du produit mixte coïncide avec celui de $\cos\theta$, et par suite, le produit mixte est positif lorsque les vecteurs **a** et $[\mathbf{b}, \mathbf{c}]$ sont orientés de la même façon par rapport au plan des vecteurs **b** et **c**, c'est-à-dire quand le triplet $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ est direct. On démontre de façon analogue que le produit mixte du triplet rétrograde de vecteurs est négatif.

Si $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ est une base orthonormée, on a $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 1$ ou $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = -1$ suivant que la base est directe ou rétrograde.

PROPOSITION 5. *Le produit mixte est nul si et seulement si les facteurs sont coplanaires.*

En effet, $(\mathbf{a}, \mathbf{b}, \mathbf{c}) = |\mathbf{a}||[\mathbf{b}, \mathbf{c}]|\cos\theta$, où θ est l'angle des vecteurs **a** et $[\mathbf{b}, \mathbf{c}]$. L'égalité $|\mathbf{a}||[\mathbf{b}, \mathbf{c}]|\cos\theta = 0$ n'est possible que si est satisfaite l'une au moins des conditions :

a) $|\mathbf{a}| = 0$. Alors les vecteurs sont évidemment coplanaires.

b) $|[\mathbf{b}, \mathbf{c}]| = 0$. Alors **b** et **c** sont colinéaires et, partant, **a**, **b** et **c** sont coplanaires.

c) $\cos\theta = 0$. Alors le vecteur **a** est orthogonal à $[\mathbf{b}, \mathbf{c}]$, c'est-à-dire est coplanaire à **b** et **c**.

La proposition réciproque se démontre de façon analogue : si **a**, **b** et **c** sont coplanaires et on n'est pas dans les cas a) et b), c'est le cas c) qui joue.

Avec la permutation des facteurs dans le produit mixte, seule l'orientation du triplet des vecteurs peut au plus subir une variation. Aussi, en vertu des propositions 4 et 5, seul le signe du produit mixte peut-il se modifier. Pour tous vecteurs **a**, **b** et **c**, on obtient en comparant les orientations des

triplets de vecteurs :

$$(\mathbf{a}, \mathbf{b}, \mathbf{c}) = (\mathbf{c}, \mathbf{a}, \mathbf{b}) = (\mathbf{b}, \mathbf{c}, \mathbf{a}) = \\ = -(\mathbf{b}, \mathbf{a}, \mathbf{c}) = -(\mathbf{c}, \mathbf{b}, \mathbf{a}) = -(\mathbf{a}, \mathbf{c}, \mathbf{b}). \quad (6)$$

En appliquant la proposition 2 au produit scalaire $(\lambda\mathbf{a}_1 + \mu\mathbf{a}_2, [\mathbf{b}, \mathbf{c}])$, on obtient l'identité

$$(\lambda\mathbf{a}_1 + \mu\mathbf{a}_2, \mathbf{b}, \mathbf{c}) = \lambda(\mathbf{a}_1, \mathbf{b}, \mathbf{c}) + \mu(\mathbf{a}_2, \mathbf{b}, \mathbf{c}). \quad (7)$$

Les égalités (6) entraînent des identités analogues pour les autres facteurs. Par exemple, pour le deuxième facteur on a

$$(\mathbf{a}, \lambda\mathbf{b}_1 + \mu\mathbf{b}_2, \mathbf{c}) = \lambda(\mathbf{a}, \mathbf{b}_1, \mathbf{c}) + \mu(\mathbf{a}, \mathbf{b}_2, \mathbf{c}).$$

En effet, on est en mesure de permuter les facteurs et mettre à la première place celui qui nous intéresse, chasser les parenthèses, puis procéder à la permutation inverse.

Les identités obtenues expriment la propriété de linéarité du produit mixte.

Démontrons maintenant la linéarité du produit vectoriel.

PROPOSITION 6. *Pour tous vecteurs* $\mathbf{a}$, $\mathbf{b}$ *et* $\mathbf{c}$ *et tous nombres* λ *et* μ *on a l'égalité*

$$[\lambda\mathbf{a} + \mu\mathbf{b}, \mathbf{c}] = \lambda[\mathbf{a}, \mathbf{c}] + \mu[\mathbf{b}, \mathbf{c}].$$

Pour le démontrer, profitons de la linéarité du produit mixte par rapport au deuxième facteur :

$$(\mathbf{d}, [\lambda\mathbf{a} + \mu\mathbf{b}, \mathbf{c}]) = \lambda(\mathbf{d}, [\mathbf{a}, \mathbf{c}]) + \mu(\mathbf{d}, [\mathbf{b}, \mathbf{c}]).$$

Cette égalité est vérifiée pour tout $\mathbf{d}$. On peut choisir une base orthonormée $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ et substituer successivement à $\mathbf{d}$ chacun des vecteurs de cette base. En vertu de la proposition 1, on obtient l'égalité de toutes les composantes des vecteurs $[\lambda\mathbf{a} + \mu\mathbf{b}, \mathbf{c}]$ et $\lambda[\mathbf{a}, \mathbf{c}] + \mu[\mathbf{b}, \mathbf{c}]$, et, partant, l'égalité qu'il fallait démontrer.

De façon analogue, on peut démontrer la linéarité du produit vectoriel par rapport au deuxième facteur.

5. Produits mixte et vectoriel exprimés en fonction des composantes des facteurs. Etant donné la décomposition des vecteurs $\mathbf{a}$ et $\mathbf{b}$ suivant les vecteurs de base $\mathbf{e}_1$, $\mathbf{e}_2$ et $\mathbf{e}_3$, on peut écrire, en vertu de la proposition 6,

$$[\mathbf{a}, \mathbf{b}] = [\{\alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \alpha_3\mathbf{e}_3\}, \{\beta_1\mathbf{e}_1 + \beta_2\mathbf{e}_2 + \beta_3\mathbf{e}_3\}] = \\ = (\alpha_1\beta_2 - \alpha_2\beta_1)[\mathbf{e}_1, \mathbf{e}_2] + (\alpha_2\beta_3 - \alpha_3\beta_2)[\mathbf{e}_2, \mathbf{e}_3] + \\ + (\alpha_3\beta_1 - \alpha_1\beta_3)[\mathbf{e}_3, \mathbf{e}_1]. \quad (8)$$

Comme on l'a vu dans l'exemple 1, on a dans une base orthonormée :

$$[\mathbf{e}_1, \mathbf{e}_2] = \pm\mathbf{e}_3, \ [\mathbf{e}_2, \mathbf{e}_3] = \pm\mathbf{e}_1, \ [\mathbf{e}_3, \mathbf{e}_1] = \pm\mathbf{e}_2,$$

où le signe est plus si la base $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ est directe, et moins si la base est rétrograde. Pour éviter les permanentes remarques sur l'orientation de la base, on admettra que la base est toujours directe. Ainsi donc, on obtient le théorème suivant.

THÉORÈME 2. *Dans une base orthonormée, le produit vectoriel est exprimé au moyen des composantes des facteurs par la formule*

$$[\mathbf{a}, \mathbf{b}] = (\alpha_2 \beta_3 - \alpha_3 \beta_2)\,\mathbf{e}_1 + (\alpha_3 \beta_1 - \alpha_1 \beta_3)\,\mathbf{e}_2 + (\alpha_1 \beta_2 - \alpha_2 \beta_1)\,\mathbf{e}_3.$$

(*Si la base est rétrograde, il faut mettre le signe moins devant l'un des membres de cette égalité.*)

On est maintenant en mesure de démontrer le théorème suivant.

THÉORÈME 3. *Le produit mixte des vecteurs* **a**, **b** *et* **c** *s'exprime en fonction de leurs composantes* $(\alpha_1, \alpha_2, \alpha_3)$, $(\beta_1, \beta_2, \beta_3)$ *et* $(\gamma_1, \gamma_2, \gamma_3)$ *rapportées à une base quelconque* $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ *par la formule*

$$(\mathbf{a}, \mathbf{b}, \mathbf{c}) = (\alpha_1 \beta_2 \gamma_3 + \alpha_2 \beta_3 \gamma_1 + \alpha_3 \beta_1 \gamma_2 - \\ - \alpha_3 \beta_2 \gamma_1 - \alpha_2 \beta_1 \gamma_3 - \alpha_1 \beta_3 \gamma_2)(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3).$$

Pour le démontrer, explicitons le produit $[\mathbf{b}, \mathbf{c}]$ suivant la formule (8) :

$$[\mathbf{b}, \mathbf{c}] = (\beta_2 \gamma_3 - \beta_3 \gamma_2)[\mathbf{e}_2, \mathbf{e}_3] + (\beta_3 \gamma_1 - \beta_1 \gamma_3)[\mathbf{e}_3, \mathbf{e}_1] + \\ + (\beta_1 \gamma_2 - \beta_2 \gamma_1)[\mathbf{e}_1, \mathbf{e}_2].$$

En multipliant scalairement les deux membres de cette égalité par le vecteur $\mathbf{a} = \alpha_1 \mathbf{e}_1 + \alpha_2 \mathbf{e}_2 + \alpha_3 \mathbf{e}_3$, on obtient

$$(\mathbf{a}, [\mathbf{b}, \mathbf{c}]) = \alpha_1 (\beta_2 \gamma_3 - \beta_3 \gamma_2)(\mathbf{e}_1, [\mathbf{e}_2, \mathbf{e}_3]) + \alpha_2 (\beta_3 \gamma_1 - \beta_1 \gamma_3)(\mathbf{e}_2, [\mathbf{e}_3, \mathbf{e}_1]) + \\ + \alpha_3 (\beta_1 \gamma_2 - \beta_2 \gamma_1)(\mathbf{e}_3, [\mathbf{e}_1, \mathbf{e}_2]).$$

(Les termes contenant les produits mixtes avec facteurs égaux ne sont pas écrits, vu qu'ils sont nuls.) D'où, compte tenu des égalités (6) et en réduisant les termes semblables, on obtient le résultat cherché.

6. Déterminants d'ordre 2 et 3. Les formules trouvées pour les produits vectoriel et mixte sont assez encombrantes. Pour une écriture plus parlante on utilise les déterminants d'ordre 2 et 3.

Considérons quatre nombres α_1, α_2, β_1, β_2. On peut en composer un tableau

$$\begin{Vmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{Vmatrix}$$

appelé *matrice d'ordre* 2. Le nombre $\alpha_1 \beta_2 - \alpha_2 \beta_1$ est appelé *déterminant de la matrice d'ordre* 2 ou tout simplement *déterminant d'ordre* 2. Il est

noté

$$\begin{vmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{vmatrix} .$$

Le produit vectoriel rapporté à une base orthonormée peut maintenant s'écrire comme suit :

$$[\mathbf{a}, \mathbf{b}] = \begin{vmatrix} \alpha_2 & \alpha_3 \\ \beta_2 & \beta_3 \end{vmatrix} \mathbf{e}_1 + \begin{vmatrix} \alpha_3 & \alpha_1 \\ \beta_3 & \beta_1 \end{vmatrix} \mathbf{e}_2 + \begin{vmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{vmatrix} \mathbf{e}_3 .$$

A partir des composantes de trois vecteurs on peut composer un· tableau, appelé *matrice d'ordre* 3

$$\left\| \begin{matrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{matrix} \right\| .$$

Le nombre

$$\alpha_1 \begin{vmatrix} \beta_2 & \beta_3 \\ \gamma_2 & \gamma_3 \end{vmatrix} + \alpha_2 \begin{vmatrix} \beta_3 & \beta_1 \\ \gamma_3 & \gamma_1 \end{vmatrix} + \alpha_3 \begin{vmatrix} \beta_1 & \beta_2 \\ \gamma_1 & \gamma_2 \end{vmatrix}$$

ou, ce qui revient au même,

$$\alpha_1 \begin{vmatrix} \beta_2 & \beta_3 \\ \gamma_2 & \gamma_3 \end{vmatrix} - \alpha_2 \begin{vmatrix} \beta_1 & \beta_3 \\ \gamma_1 & \gamma_3 \end{vmatrix} + \alpha_3 \begin{vmatrix} \beta_1 & \beta_2 \\ \gamma_1 & \gamma_2 \end{vmatrix}$$

est appelé *déterminant de la matrice d'ordre* 3 ou *déterminant d'ordre* 3 et noté

$$\begin{vmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix} .$$

Suivant le théorème 3, on a en nouvelles notations :

$$(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \begin{vmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix} (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3). \tag{9}$$

En particulier, on a par rapport à une base orthonormée :

$$(\mathbf{a}, \mathbf{b}, \mathbf{c}) = \begin{vmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix} . \tag{10}$$

Le théorème 2 et la définition du déterminant d'ordre 3 permettent obtenir l'écriture suivante du produit vectoriel en fonction des composan-

tes des facteurs par rapport à une base orthonormée $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$:

$$[\mathbf{a}, \mathbf{b}] = \left\| \begin{matrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \\ \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \end{matrix} \right\|. \qquad (11)$$

Les déterminants sont étroitement liés aux systèmes d'équations linéaires dont il est commode d'écrire les solutions en s'en servant. Une grande partie du chapitre V y sera consacrée. En attendant, on ne fournira qu'une illustration géométrique.

Soit donné un système de trois équations à trois inconnues

$$\left.\begin{aligned} a_1 x + b_1 y + c_1 z &= f_1, \\ a_2 x + b_2 y + c_2 z &= f_2, \\ a_3 x + b_3 y + c_3 z &= f_3. \end{aligned}\right\}$$

Fixons dans l'espace une base et considérons les vecteurs $\mathbf{a}(a_1, a_2, a_3)$, $\mathbf{b}(b_1, b_2, b_3)$, $\mathbf{c}(c_1, c_2, c_3)$ et $\mathbf{f}(f_1, f_2, f_3)$. On peut alors considérer le système comme une écriture en coordonnées de l'égalité vectorielle

$$x\mathbf{a} + y\mathbf{b} + z\mathbf{c} = \mathbf{f}. \qquad (12)$$

Les coefficients x, y, z de la décomposition de $\mathbf{f}$ suivant les vecteurs $\mathbf{a}$, $\mathbf{b}$ et $\mathbf{c}$ représentent donc une solution du système. On peut s'assurer que le système possède une solution unique si $\mathbf{a}$, $\mathbf{b}$ et $\mathbf{c}$ ne sont pas coplanaires, c'est-à-dire si $(\mathbf{a}, \mathbf{b}, \mathbf{c}) \neq 0$. Admettons que cette condition est satisfaite et cherchons la solution. A cet effet, multiplions les deux membres de l'égalité (12) scalairement par le produit vectoriel $[\mathbf{b}, \mathbf{c}]$. On obtient $x(\mathbf{a}, \mathbf{b}, \mathbf{c}) =$ $= (\mathbf{f}, \mathbf{b}, \mathbf{c})$ et, par suite, x est égal au rapport des déterminants

$$\begin{vmatrix} f_1 & f_2 & f_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix} \quad \text{et} \quad \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}.$$

De façon analogue on obtient les autres inconnues.

Arrêtons-nous sur les propriétés suivantes des déterminants. Il découle des égalités (6) que le déterminant change de signe lorsqu'on permute deux lignes de la matrice. La formule (7) nous donne

$$\begin{vmatrix} \lambda\alpha_1 + \mu\alpha_1' & \lambda\alpha_2 + \mu\alpha_2' & \lambda\alpha_3 + \mu\alpha_3' \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix} = \lambda \begin{vmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix} + \mu \begin{vmatrix} \alpha_1' & \alpha_2' & \alpha_3' \\ \beta_1 & \beta_2 & \beta_3 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{vmatrix}.$$

7. Conditions de colinéarité et de coplanarité des vecteurs. Commençons par énoncer la proposition utile suivante.

PROPOSITION 7. *Quelle que soit la base* $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, *les produits vectoriels* $[\mathbf{e}_1, \mathbf{e}_2]$, $[\mathbf{e}_2, \mathbf{e}_3]$, $[\mathbf{e}_3, \mathbf{e}_1]$ *sont linéairement indépendants.*

Pour la démontrer, raisonnons par l'absurde. Considérons une combinaison linéaire des vecteurs qui nous intéressent et supposons qu'elle soit égale à zéro :

$$\lambda[\mathbf{e}_2, \mathbf{e}_3] + \mu[\mathbf{e}_3, \mathbf{e}_1] + \nu[\mathbf{e}_1, \mathbf{e}_2] = 0.$$

Admettons qu'un des coefficients, soit λ pour fixer les idées, est différent de zéro. Multiplions la combinaison linéaire par $\mathbf{e}_1$. On obtient $\lambda(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 0$, c'est-à-dire $\lambda = 0$. La contradiction obtenue démontre la proposition.

Les propositions qui suivent énoncent les conditions nécessaires et suffisantes de la colinéarité et de la coplanarité des vecteurs.

PROPOSITION 8. *Pour que trois vecteurs soient coplanaires il est nécessaire et suffisant que la matrice de leurs composantes ait le déterminant nul.*

La proposition 8 résulte immédiatement de la proposition 5 et de la formule (9), vu que $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) \neq 0$.

PROPOSITION 9. *Soient* $(\alpha_1, \alpha_2, \alpha_3)$ *et* $(\beta_1, \beta_2, \beta_3)$ *les composantes des vecteurs* $\mathbf{a}$ *et* $\mathbf{b}$ *par rapport à une base donnée quelconque. Ces vecteurs sont colinéaires si et seulement si*

$$\begin{vmatrix} \alpha_2 & \alpha_3 \\ \beta_2 & \beta_3 \end{vmatrix} = \begin{vmatrix} \alpha_3 & \alpha_1 \\ \beta_3 & \beta_1 \end{vmatrix} = \begin{vmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{vmatrix} = 0. \qquad (13)$$

La suffisance de la condition est évidente : de l'égalité (13), en appliquant la formule (8), on obtient que $[\mathbf{a}, \mathbf{b}] = 0$, ce qui équivaut à la colinéarité des vecteurs. Notons que la formule (8) est vérifiée par rapport à toute base, de sorte qu'on peut ne pas exiger qu'elle soit orthonormée. Inversement, de la condition $[\mathbf{a}, \mathbf{b}] = 0$ et de la formule (8) on peut déduire (13), car les vecteurs $[\mathbf{e}_2, \mathbf{e}_3]$, $[\mathbf{e}_3, \mathbf{e}_1]$ et $[\mathbf{e}_1, \mathbf{e}_2]$ sont linéairement indépendants en vertu de la proposition 7.

En géométrie plane, la colinéarité de deux vecteurs s'établit par la proposition suivante.

PROPOSITION 10. *Pour que deux vecteurs dans le plan soient colinéaires il est nécessaire et suffisant que la matrice de leurs composantes ait le déterminant nul.*

Pour le démontrer, on admettra que le plan considéré est placé dans l'espace et que la base dans le plan est complétée par le troisième vecteur pour avoir une base dans l'espace. Le vecteur $\mathbf{a}$ de coordonnées (α_1, α_2) dans le plan acquiert alors les coordonnées $(\alpha_1, \alpha_2, 0)$ relativement à la

base dans l'espace. En appliquant maintenant la proposition 9 aux vecteurs $\mathbf{a}(\alpha_1, \alpha_2)$ et $\mathbf{b}(\beta_1, \beta_2)$ on a une seule condition

$$\begin{vmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{vmatrix} = 0$$

de leur colinéarité (les deux autres déterminants sont nuls puisque $\alpha_3 = \beta_3 = 0$).

8. Aire d'un parallélogramme. Etant donné dans l'espace deux vecteurs non colinéaires **a** et **b** d'origine commune, l'aire du parallélogramme construit sur ces vecteurs s'exprime en fonction de leurs composantes par rapport à la base orthonormée suivant la formule

$$S = |[\mathbf{a}, \mathbf{b}]| = \sqrt{(\alpha_2 \beta_3 - \alpha_3 \beta_2)^2 + (\alpha_3 \beta_1 - \alpha_1 \beta_3)^2 + (\alpha_1 \beta_2 - \alpha_2 \beta_1)^2}. \quad (14)$$

En géométrie plane, l'aire d'un parallélogramme se calcule de façon analogue. Bien que le produit vectoriel ne soit pas défini, on peut, comme dans la démonstration de la proposition 10, admettre que le plan étudié est placé dans l'espace et que pour troisième vecteur de base est choisi un vecteur de longueur unité perpendiculaire à ce plan. Dans ce cas, le produit vectoriel de deux vecteurs dans le plan n'a qu'une composante différente de zéro, à savoir la troisième, et l'aire du parallélogramme construit sur les vecteurs **a** et **b** s'exprime au moyen de leurs composantes, par rapport à la base orthonormée du plan, par la formule

$$S = |\alpha_1 \beta_2 - \alpha_2 \beta_1|, \quad (15)$$

autrement dit, est égale à la valeur absolue du déterminant

$$\begin{vmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{vmatrix}.$$

Quand on écrit l'expression $\alpha_1\beta_2 - \alpha_2\beta_1$, il est essentiel d'avoir en vue l'ordre des vecteurs, c'est-à-dire que **a** est le premier et **b** est le second. Si **a** était le second et **b** le premier, la composante du produit vectoriel serait $\beta_1\alpha_2 - \beta_2\alpha_1$. L'aire du parallélogramme est indépendante de l'ordre des vecteurs, car elle est égale au module de cette expression.

Introduisons maintenant la notion d'orientation du parallélogramme dans le plan. On supposera que le parallélogramme est construit sur deux vecteurs, c'est-à-dire que ses deux côtés adjacents sont des vecteurs d'origine commune.

DÉFINITION. Un parallélogramme est dit *orienté* si le couple de vecteurs sur lesquels il est construit est ordonné. L'orientation est dite *positive* si la plus petite rotation appliquant le premier vecteur sur le second s'effectue dans le sens inverse au mouvement des aiguilles d'une montre, et *négative* dans le cas contraire.

On supposera que le parallélogramme construit sur les vecteurs de base $\mathbf{e}_1$ et $\mathbf{e}_2$ (parallélogramme de base) est orienté positivement. On choisit pour $\mathbf{e}_3$ le vecteur $[\mathbf{e}_1, \mathbf{e}_2]$ vu que dans l'espace une base est toujours directe. L'inégalité $\alpha_1\beta_2 - \alpha_2\beta_1 > 0$ signifie que le vecteur

$[\mathbf{a}, \mathbf{b}]$ est orienté de la même façon que $[\mathbf{e}_1, \mathbf{e}_2]$ et, partant, le parallélogramme construit sur les vecteurs $\mathbf{a}$ et $\mathbf{b}$ a aussi une orientation positive.

On convient de représenter l'aire d'un parallélogramme orienté par un nombre affecté d'un signe : positif pour les parallélogrammes orientés positivement, et négatif pour les parallélogrammes orientés négativement.

Comme le montrent les raisonnements précédents, si le parallélogramme de base est orienté positivement, le nombre $\alpha_1\beta_2 - \alpha_2\beta_1$ est égal à l'aire du parallélogramme orienté construit sur les vecteurs $\mathbf{a}(\alpha_1, \alpha_2)$ et $\mathbf{b}(\beta_1, \beta_2)$. On peut aussi dire que l'orientation du parallélogramme construit sur les vecteurs $\mathbf{a}$ et $\mathbf{b}$ coïncide avec celle du parallélogramme de base si $\alpha_1\beta_2 - \alpha_2\beta_1 > 0$ et qu'elle est contraire si $\alpha_1\beta_2 - \alpha_2\beta_1 < 0$.

9. Volume d'un parallélépipède orienté. On supposera que le parallélépipède est construit sur trois vecteurs, c'est-à-dire que ses trois arêtes sont des vecteurs ayant une origine commune.

DÉFINITION. Un parallélépipède est dit *orienté* si le triplet de vecteurs sur lesquels il est construit est ordonné. L'orientation est dite *positive* si le triplet est direct et *négative* dans le cas contraire.

On admet que le volume d'un parallélépipède orienté est positif si son orientation est positive, et négatif si elle est négative. La proposition 4 peut maintenant être formulée de la façon suivante.

PROPOSITION 11. *Le produit mixte de trois vecteurs non coplanaires est égal au volume d'un parallélépipède orienté construit sur ces vecteurs.*

10. Produit vectoriel double. L'expression $[\mathbf{a}, [\mathbf{b}, \mathbf{c}]]$ est appelée *produit vectoriel double* des vecteurs $\mathbf{a}, \mathbf{b}, \mathbf{c}$. Démontrons que

$$[\mathbf{a}, [\mathbf{b}, \mathbf{c}]] = (\mathbf{a}, \mathbf{c})\,\mathbf{b} - (\mathbf{a}, \mathbf{b})\,\mathbf{c}. \tag{16}$$

A cet effet, choisissons une base orthonormée $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ de manière que le vecteur $\mathbf{e}_1$ soit de même sens que $\mathbf{a}$ (c'est-à-dire que $\mathbf{a} = \alpha\mathbf{e}_1$, où $\alpha = |\mathbf{a}|$). Notons $(\beta_1, \beta_2, \beta_3)$, $(\gamma_1, \gamma_2, \gamma_3)$ et $(\Delta_1, \Delta_2, \Delta_3)$ les coordonnées respectives des vecteurs $\mathbf{b}$, $\mathbf{c}$ et $[\mathbf{b}, \mathbf{c}]$. Selon la formule (1) on a alors

$$(\mathbf{a}, \mathbf{c})\,\mathbf{b} = \alpha\gamma_1\,(\beta_1\,\mathbf{e}_1 + \beta_2\,\mathbf{e}_2 + \beta_3\,\mathbf{e}_3)$$

et

$$(\mathbf{a}, \mathbf{b})\,\mathbf{c} = \alpha\beta_1\,(\gamma_1\,\mathbf{e}_1 + \gamma_2\,\mathbf{e}_2 + \gamma_3\,\mathbf{e}_3).$$

Cela permet de transformer le second membre de la formule (16) :

$$\alpha(\gamma_1\beta_2 - \beta_1\gamma_2)\,\mathbf{e}_2 + \alpha(\gamma_1\beta_3 - \beta_1\gamma_3)\,\mathbf{e}_3 = -\alpha\Delta_3\,\mathbf{e}_2 + \alpha\Delta_2\,\mathbf{e}_3.$$

D'autre part, de la formule (11) on déduit immédiatement que

$$[\mathbf{a}, [\mathbf{b}, \mathbf{c}]] = -\alpha\Delta_3\,\mathbf{e}_2 + \alpha\Delta_2\,\mathbf{e}_3,$$

ce qui achève la démonstration.

11. Base biorthogonale. DÉFINITION. Une base composée des vecteurs

$$\mathbf{e}_1^* = \frac{[\mathbf{e}_2, \mathbf{e}_3]}{(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)}, \quad \mathbf{e}_2^* = \frac{[\mathbf{e}_3, \mathbf{e}_1]}{(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)}, \quad \mathbf{e}_3^* = \frac{[\mathbf{e}_1, \mathbf{e}_2]}{(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)}$$

est dite *biorthogonale* à la base $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

La proposition 7 implique que $\mathbf{e}_1^*$, $\mathbf{e}_2^*$, $\mathbf{e}_3^*$ sont non coplanaires et constituent de fait une base.

Il n'est pas difficile de vérifier que toute base orthonormée se confond avec la base biorthogonale.

PROPOSITION 12. *Si* $\{\mathbf{e}_1^*, \mathbf{e}_2^*, \mathbf{e}_3^*\}$ *est la base biorthogonale à la base* $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, *les composantes d'un vecteur arbitraire* $\mathbf{a}$ *par rapport à* $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ *se calculent par les formules*

$$\alpha_1 = (\mathbf{a}, \mathbf{e}_1^*), \quad \alpha_2 = (\mathbf{a}, \mathbf{e}_2^*), \quad \alpha_3 = (\mathbf{a}, \mathbf{e}_3^*).$$

En effet, en multipliant l'égalité $\mathbf{a} = \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \alpha_3\mathbf{e}_3$ scalairement par $\mathbf{e}_1^*$, on obtient $(\mathbf{a}, \mathbf{e}_1^*) = \alpha_1 \dfrac{(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)}{(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)}$. De façon analogue on démontre les autres formules.

PROPOSITION 13. *Si* $\{\mathbf{e}_1^*, \mathbf{e}_2^*, \mathbf{e}_3^*\}$ *est la base biorthogonale à* $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, *la base* $\{\mathbf{e}_1^{**}, \mathbf{e}_2^{**}, \mathbf{e}_3^{**}\}$ *qui est biorthogonale à* $\{\mathbf{e}_1^*, \mathbf{e}_2^*, \mathbf{e}_3^*\}$ *se confond avec* $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

Pour le démontrer, commençons par décomposer le vecteur $\mathbf{e}_1^*$ suivant les vecteurs $\mathbf{e}_1^*$, $\mathbf{e}_2^*$, $\mathbf{e}_3^*$ en utilisant les formules de la proposition 12. Vu que les composantes de $\mathbf{e}_1^*$ sont (1, 0, 0), il vient

$$(\mathbf{e}_1^*, \mathbf{e}_1^{**}) = 1, \quad (\mathbf{e}_1^*, \mathbf{e}_2^{**}) = (\mathbf{e}_1^*, \mathbf{e}_3^{**}) = 0.$$

En décomposant de façon analogue $\mathbf{e}_2^*$ et $\mathbf{e}_3^*$ on obtient encore six égalités

$$(\mathbf{e}_2^*, \mathbf{e}_2^{**}) = 1, \quad (\mathbf{e}_2^*, \mathbf{e}_1^{**}) = (\mathbf{e}_2^*, \mathbf{e}_3^{**}) = 0,$$
$$(\mathbf{e}_3^*, \mathbf{e}_3^{**}) = 1, \quad (\mathbf{e}_3^*, \mathbf{e}_1^{**}) = (\mathbf{e}_3^*, \mathbf{e}_2^{**}) = 0.$$

Or ces égalités signifient en vertu de la même proposition 12 que les composantes des vecteurs $\mathbf{e}_1^{**}$, $\mathbf{e}_2^{**}$ et $\mathbf{e}_3^{**}$ par rapport à la base $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ sont respectivement égales à (1, 0, 0), (0, 1, 0) et (0, 0, 1), c'est-à-dire que $\mathbf{e}_1^{**} = \mathbf{e}_1$, $\mathbf{e}_2^{**} = \mathbf{e}_2$, $\mathbf{e}_3^{**} = \mathbf{e}_3$. La proposition est démontrée.

12. A propos des grandeurs vectorielles. Dans les mathématiques appliquées, on étudie souvent des grandeurs représentées par les vecteurs : forces, vitesses, moments de forces, etc. Chacune de ces grandeurs vectorielles peut être définie par une formule de dimensions. Sans entrer dans le vif du sujet, on énoncera seulement les règles formelles d'opérations sur ces formules.

Du point de vue formel, la formule de dimensions de la grandeur est un monôme composé d'un ensemble de symboles. On multiplie et divise ces monômes suivant les règles habituelles des opérations algébriques sur les mônomes :

1. On n'additionne les grandeurs vectorielles que si leurs formules de dimensions coïncident, la formule de dimensions de la somme étant celle de ses termes.

2. En multipliant une grandeur vectorielle par une grandeur scalaire, on multiplie leurs formules de dimensions.

3. La grandeur vectorielle et son module ont une même formule de dimensions.

4. La formule de dimensions du produit scalaire (du produit vectoriel)

est égale au produit des formules de dimensions des facteurs, ce qui découle aisément de leurs définitions et de la règle précédente.

Pour représenter une grandeur vectorielle sur le dessin on doit se mettre d'accord sur l'échelle à utiliser : combien d'unités de longueur (par exemple, de cm) il faut prendre pour représenter une unité de grandeur ayant la formule de dimensions donnée (par exemple, km, m/s, N).

Si les facteurs du produit vectoriel sont mesurés en unités de longueur, le produit est mesuré en unités d'aire. Pour représenter les unités d'aire on choisit l'échelle de manière qu'une unité d'aire soit représentée par une unité linéaire. Dans ce cas, la longueur du produit vectoriel sera numériquement égale à l'aire du parallélogramme construit sur les facteurs.

Etant donné que l'unité de longueur est choisie une fois pour toutes et ne varie pas, cette convention n'entraîne pas des contradictions. Mais le problème n'est pas aussi bénin qu'il paraît. C'est ainsi que deux mathématiciens utilisant cette convention mais se servant d'unités de longueur différentes (un mathématicien français utilisant les centimètres et son collègue anglais, les pouces) dessineront, pour les mêmes vecteurs, des produits vectoriels qui ne coïncideront pas.

§ 4. Changement de base et de repère

1. Changement de base. Jusqu'à présent nous avons supposé qu'il existait une seule base. Mais rien ne limite le choix d'une base, et un problème important se présente : rechercher les composantes d'un vecteur par rapport à une base donnée à partir de ses composantes dans une autre base. Ceci étant, on doit préciser la position de la nouvelle base par rapport à l'ancienne, à savoir, on doit donner les composantes des nouveaux vecteurs de base $\mathbf{e}_1'$, $\mathbf{e}_2'$, $\mathbf{e}_3'$ par rapport à l'ancienne base $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$. Soit *)

$$\left.\begin{aligned} \mathbf{e}_1' &= a_1^1\,\mathbf{e}_1 + a_1^2\,\mathbf{e}_2 + a_1^3\,\mathbf{e}_3, \\ \mathbf{e}_2' &= a_2^1\,\mathbf{e}_1 + a_2^2\,\mathbf{e}_2 + a_2^3\,\mathbf{e}_3, \\ \mathbf{e}_3' &= a_3^1\,\mathbf{e}_1 + a_3^2\,\mathbf{e}_2 + a_3^3\,\mathbf{e}_3. \end{aligned}\right\} \qquad (1)$$

Décomposons un vecteur arbitraire $\mathbf{a}$ suivant les vecteurs $\mathbf{e}_1'$, $\mathbf{e}_2'$, $\mathbf{e}_3'$:

$$\mathbf{a} = \alpha_1'\,\mathbf{e}_1' + \alpha_2'\,\mathbf{e}_2' + \alpha_3'\,\mathbf{e}_3'.$$

Notons α_1, α_2 et α_3 les composantes du vecteur $\mathbf{a}$ par rapport à l'ancienne base. En décomposant chaque terme de l'égalité précédente suivant les vec-

*) Pour commodité, un des indices est placé en haut. Ce n'est pas un exposant. Par exemple, a_3^1 se lit « a un-trois ».

teurs $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ on obtient en vertu des propositions 4 et 3 du § 1 :

$$\left.\begin{aligned} \alpha_1 &= a_1^1\alpha_1' + a_2^1\alpha_2' + a_3^1\alpha_3', \\ \alpha_2 &= a_1^2\alpha_1' + a_2^2\alpha_2' + a_3^2\alpha_3', \\ \alpha_3 &= a_1^3\alpha_1' + a_2^3\alpha_2' + a_3^3\alpha_3'. \end{aligned}\right\} \qquad (2)$$

Les relations (2) définissent la solution du problème posé. Si on a besoin d'exprimer les nouvelles composantes en fonction des composantes anciennes, il faudra résoudre le système d'équations (2) par rapport aux inconnues $\alpha_1', \alpha_2', \alpha_3'$.

De la même façon, on peut obtenir les formules liant les composantes d'un vecteur par rapport à différentes bases du plan. Ces formules sont :

$$\left.\begin{aligned} \alpha_1 &= a_1^1\alpha_1' + a_2^1\alpha_2', \\ \alpha_2 &= a_1^2\alpha_1' + a_2^2\alpha_2'. \end{aligned}\right\} \qquad (2')$$

Les coefficients associés à α_1', α_2' et α_3' dans les formules (2) peuvent être écrits sous la forme d'une matrice d'ordre 3 :

$$\begin{Vmatrix} a_1^1 & a_2^1 & a_3^1 \\ a_1^2 & a_2^2 & a_3^2 \\ a_1^3 & a_2^3 & a_3^3 \end{Vmatrix}. \qquad (3)$$

Cette matrice porte le nom de *matrice de passage* de la base $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ à la base $\{\mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$. Dans ses colonnes se trouvent les composantes des nouveaux vecteurs de base $\mathbf{e}_1'$, $\mathbf{e}_2'$ et $\mathbf{e}_3'$ rapportés à l'ancienne base $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

2. Changement de repère. Considérons maintenant deux repères cartésiens : l'ancien $\{O, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ et le nouveau $\{O', \mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$. Soit M un point quelconque dont les coordonnées par rapport à ces repères sont notées respectivement (x, y, z) et (x', y', z'). On se propose d'exprimer x, y, z en fonction de x', y', z' en supposant connue la position du nouveau repère par rapport à l'ancien, autrement dit connues les anciennes coordonnées a_0^1, a_0^2, a_0^3 de la nouvelle origine O', et les composantes des nouveaux vecteurs de base par rapport à l'ancienne base, qui constituent la matrice de passage (3).

Les rayons vecteurs du point M relativement aux points O et O' sont liés par l'égalité $\overrightarrow{OM} = \overrightarrow{OO'} + \overrightarrow{O'M}$ qu'on peut écrire sous la forme

$$\overrightarrow{OM} = \overrightarrow{OO'} + x'\mathbf{e}_1' + y'\mathbf{e}_2' + z'\mathbf{e}_3', \qquad (4)$$

vu que x', y', z' sont les composantes de $\overrightarrow{O'M}$ par rapport à la base $\{\mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$. Décomposons chaque terme de l'égalité (4) suivant les vecteurs $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$, compte tenu du fait que les composantes de $\overrightarrow{OM}$ et de $\overrightarrow{OO'}$ sont égales aux coordonnées des points M et O', notées respectivement

x, y, z et a_0^1, a_0^2, a_0^3. On obtient trois égalités numériques équivalentes à l'égalité (4) :

$$\left.\begin{aligned} x &= a_1^1 x' + a_2^1 y' + a_3^1 z' + a_0^1, \\ y &= a_1^2 x' + a_2^2 y' + a_3^2 z' + a_0^2, \\ z &= a_1^3 x' + a_2^3 y' + a_3^3 z' + a_0^3. \end{aligned}\right\} \qquad (5)$$

Les égalités (5) déterminent justement la transformation des coordonnées d'un point lors du passage d'un repère cartésien à l'autre.

3. Transformation d'un repère cartésien rectangulaire dans le plan. Dans le plan, les formules de passage d'un repère à un autre peuvent être obtenues à partir de (5) si l'on y conserve les deux premières égalités en omettant les termes contenant z :

$$\left.\begin{aligned} x &= a_1^1 x' + a_2^1 y' + a_0^1, \\ y &= a_1^2 x' + a_2^2 y' + a_0^2. \end{aligned}\right\} \qquad (6)$$

Considérons un cas particulier de deux repères cartésiens rectangulaires.

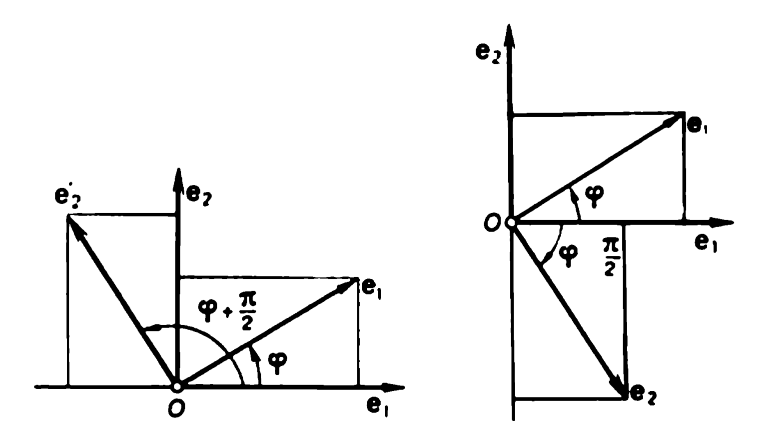


Fig. 14. Deux cas de position relative de deux bases orthonormées dans le plan

Désignons par φ l'angle des vecteurs $\mathbf{e}_1$ et $\mathbf{e}_1'$ mesuré dans le sens de la plus petite rotation amenant $\mathbf{e}_1$ sur $\mathbf{e}_2$. Alors (fig. 14) on a

$$\mathbf{e}_1' = \cos\varphi\,\mathbf{e}_1 + \sin\varphi\,\mathbf{e}_2,$$

$$\mathbf{e}_2' = \cos\left(\varphi \pm \frac{\pi}{2}\right)\mathbf{e}_1 + \sin\left(\varphi \pm \frac{\pi}{2}\right)\mathbf{e}_2.$$

On met le signe plus dans la décomposition de $\mathbf{e}_2'$ si la plus petite rotation de $\mathbf{e}_1'$ à $\mathbf{e}_2'$ est de même sens que celle de $\mathbf{e}_1$ à $\mathbf{e}_2$, autrement dit si la nouvelle base est obtenue de l'ancienne par rotation d'angle φ. On met le signe moins dans la décomposition de $\mathbf{e}_2'$ si au contraire la nouvelle base ne peut

être obtenue par rotation de l'ancienne. Vu que

$$\cos\left(\varphi \pm \frac{\pi}{2}\right) = \mp \sin\varphi, \quad \sin\left(\varphi \pm \frac{\pi}{2}\right) = \pm\cos\varphi,$$

on en déduit

$$\left.\begin{aligned} x &= x'\cos\varphi \mp y'\sin\varphi + a_0^1, \\ y &= x'\sin\varphi \pm y'\cos\varphi + a_0^2, \end{aligned}\right\} \tag{7}$$

où l'on prend le signe supérieur si la rotation du repère est possible.

CHAPITRE II

DROITES ET PLANS

§ 1. Notions générales sur les équations

1. Définition. Soient un repère et un ensemble S rapporté à ce repère. Par *équation de l'ensemble* S on entend une expression qui définit S au moyen des coordonnées de ses points, c'est-à-dire une assertion qui est vraie pour les coordonnées des points appartenant à S, et qui est fausse pour les coordonnées des points qui ne lui appartiennent pas.

En géométrie plane, une équation d'un ensemble de points se présente souvent sous la forme $F(x, y) = 0$, où F est une fonction de deux variables, et en géométrie dans l'espace, sous la forme $F(x, y, z) = 0$ où F est une fonction de trois variables.

En étudiant les mathématiques, on fait connaissance avec des règles logiques et mathématiques qui permettent de passer d'une assertion vraie à l'autre. L'étude rigoureuse de ces règles est du domaine d'une science spéciale, la logique mathématique. Quant à nous, en formulant les propositions qui suivront on admettra que ces règles sont connues au lecteur. Il est donc naturel qu'on s'abstienne de les démontrer.

PROPOSITION 1. *Si P_S et P_T sont les équations des ensembles S et T, l'équation de leur intersection $S \cap T$ est une assertion affirmant que P_S et P_T sont vraies en même temps.*

Soient P_S et P_T deux égalités contenant les coordonnées d'un point : $F_S(x, y, z) = 0$ et $F_T(x, y, z) = 0$. Alors l'équation de l'intersection est un système d'équations :

$$F_S(x, y, z) = 0, \quad F_T(x, y, z) = 0.$$

PROPOSITION 2. *Si P_S et P_T sont les équations des ensembles S et T, l'équation de leur réunion $S \cup T$ est une assertion affirmant que de P_S et P_T l'une au moins est vraie.*

Si P_S et P_T sont de la forme $F_S(x, y, z) = 0$ et $F_T(x, y, z) = 0$, l'équation de la réunion peut être écrite sous la forme

$$F_S(x, y, z) \cdot F_T(x, y, z) = 0.$$

PROPOSITION 3. *Si P_S et P_T sont les équations des ensembles S et T, et S est un sous-ensemble de T, P_S implique P_T.*

PROPOSITION 4. *Les ensembles S et T se confondent si et seulement si leurs équations sont équivalentes, c'est-à-dire si P_S implique P_T, et P_T implique P_S.*

Parfois les propositions 3 et 4 sont considérées comme les définitions des relations « implication » et « équivalence » entre les équations.

2. Courbes et surfaces algébriques. L'étude d'ensembles de points quelconques est un domaine immense. On définira ici une classe d'ensembles relativement étroite mais toutefois beaucoup trop large pour être étudiée de façon détaillée.

DÉFINITION. On appelle *surface algébrique* un ensemble qui dans un repère cartésien peut être défini par une équation de la forme

$$A_1 x^{k_1} y^{l_1} z^{m_1} + \ldots + A_s x^{k_s} y^{l_s} z^{m_s} = 0, \qquad (1)$$

où tous les exposants sont des entiers positifs. La plus grande des sommes *) $k_1 + l_1 + m_1, \ldots, k_s + l_s + m_s$ est appelée *degré de l'équation* ou *ordre de la surface algébrique.*

DÉFINITION. On appelle *courbe algébrique* plane un ensemble qui dans un repère cartésien du plan peut être défini par une équation de la forme

$$A_1 x^{k_1} y^{l_1} + \ldots + A_s x^{k_s} y^{l_s} = 0, \qquad (2)$$

tous les exposants étant des entiers positifs. La plus grande des sommes $k_1 + l_1, \ldots, k_s + l_s$ est appelée *degré de l'équation* ou *ordre de la courbe.*

On voit sans peine qu'une surface algébrique n'est pas obligatoirement telle qu'on se la représente intuitivement. Par exemple, l'équation $x^2 + y^2 + z^2 + 1 = 0$ n'est vérifiée par les coordonnées d'aucun point. L'équation

$$(x^2 + y^2 + z^2)[(x - 1)^2 + (y - 1)^2 + (z - 1)^2] = 0$$

définit deux points, l'équation $y^2 + z^2 = 0$ définit une droite (l'axe des abscisses). La même remarque concerne les courbes algébriques. Le lecteur trouvera lui-même des exemples appropriés.

Les définitions données possèdent un important défaut. A savoir, on ignore la forme que prend l'équation de la surface dans un autre repère cartésien. Même si dans un autre repère cartésien une équation présente la forme (1), lequel des degrés de ces équations doit être appelé ordre de la

---

*) Il s'agit évidemment de la plus grande somme figurant dans l'équation, c'est-à-dire qu'on suppose qu'après avoir réduit les termes semblables il existe au moins un terme de coefficient non nul possédant une telle somme d'exposants. La même remarque se rapporte à la définition de l'ordre de la courbe algébrique, donnée plus bas.

surface. Des questions analogues se posent aussi pour des courbes algébriques. La réponse nous est fournie par les théorèmes suivants appelés *théorèmes d'invariance de l'ordre.*

THÉORÈME 1. *Toute surface définie par une équation de la forme* (1) *dans un repère cartésien l'est aussi dans tout autre repère cartésien, le degré de l'équation restant le même.*

THÉORÈME 2. *Toute courbe du plan définie par une équation de la forme* (2) *dans un repère cartésien l'est aussi dans tout autre repère cartésien, le degré de l'équation restant le même.*

Les deux théorèmes se démontrent de la même façon. Démontrons par exemple le théorème 2. Passons du repère cartésien $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ dont il s'agissait dans la définition à un nouveau repère cartésien quelconque $\{O', \mathbf{e}_1', \mathbf{e}_2'\}$. Les anciennes coordonnées x, y sont liées aux nouvelles x', y' par les formules (6) du § 4, ch. I :

$$x = a_1^1 x' + a_2^1 y' + a_0^1,$$
$$y = a_1^2 x' + a_2^2 y' + a_0^2.$$

Pour obtenir l'équation de la courbe dans le nouveau repère, portons dans son équation les expressions de x et y en fonction de x' et y'. En élevant le trinôme $a_1^1 x' + a_2^1 y' + a_0^1$ à la puissance k on obtient un polynôme en x' et y' de degré k. En élevant $a_1^2 x' + a_2^2 y' + a_0^2$ à la puissance l on obtient un polynôme de degré l. En multipliant les polynômes obtenus on remarque que chaque terme de la forme $Ax^k y^l$ figurant dans le premier membre de l'équation (2) devient un polynôme de degré $k + l$ en x' et y'. La somme des polynômes est un polynôme dont le degré est au plus égal aux degrés de ses termes. (Il aurait pu être strictement inférieur si les termes de plus haut degré étaient supprimés.) On vient ainsi de démontrer que dans tout repère cartésien la courbe algébrique se définit par une équation de la forme (2) et que le degré de cette équation ne peut augmenter avec le passage d'un repère à un autre. Il nous reste à démontrer qu'il ne peut diminuer non plus et, par suite, doit rester constant. On le démontre aisément par l'absurde. En effet, avec le passage inverse du repère $\{O', \mathbf{e}_1', \mathbf{e}_2'\}$ au repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$, les anciennes coordonnées x', y' d'un point s'expriment en fonction de ses nouvelles coordonnées x, y au moyen des formules analogues à celles données ci-dessus. Si, avec le passage de $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ à $\{O', \mathbf{e}_1', \mathbf{e}_2'\}$, le polynôme $F(x, y)$ se transformait en polynôme $G(x', y')$, le passage inverse convertit le polynôme $G(x', y')$ en $F(x, y)$. Supposons maintenant qu'au cours du passage du repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ au repère $\{O', \mathbf{e}_1', \mathbf{e}_2'\}$ le degré de l'équation diminue. Alors, avec le passage inverse de $\{O', \mathbf{e}_1', \mathbf{e}_2'\}$ à $\{O, \mathbf{e}_1, \mathbf{e}_2\}$, le degré aurait dû augmenter, ce qui, comme on le sait, est impossible.

REMARQUE. La propriété d'invariance de l'ordre ne se rapporte pas aux différentes équations qu'une courbe ou une surface peut avoir dans un même repère. Bien que ces équations soient équivalentes, il existe parmi elles des équations de différents degrés et même de formes autres que (1) ou (2). En effet, les trois équations suivantes définissent le cercle de rayon 1 et de centre à l'origine du repère cartésien rectangulaire :

$$\sqrt{x^2 + y^2} = 1, \quad x^2 + y^2 - 1 = 0, \quad (x^2 + y^2 - 1)^2 = 0. \qquad (3)$$

On admet que les équations équivalentes de forme (2) et de degrés différents définissent des courbes algébriques différentes (bien que les ensembles de points qui les vérifient se confondent). Par exemple, on dit que la dernière équation (3) définit un « cercle double ».

L'ordre d'une courbe algébrique est le premier exemple d'invariant rencontré. D'une façon générale, on appelle *invariant* toute grandeur qui ne varie pas avec le changement de repère. Seules les combinaisons invariantes de grandeurs (de coefficients, d'exposants, etc.) intervenant dans l'équation d'une courbe ou d'une surface caractérisent les propriétés géométriques de la courbe ou de la surface indépendamment de leur position par rapport au repère. Quant à l'interprétation géométrique de l'ordre de la courbe, elle sera éclaircie à la fin de ce chapitre.

On est en mesure maintenant de préciser l'objet principal du cours de géométrie analytique. On y groupe essentiellement des courbes et des surfaces algébriques d'ordre 1 et 2 se prêtant à l'étude par les procédés de l'algèbre élémentaire.

Avant d'aborder leurs propriétés, passons en revue quelques équations plus générales. On aura affaire à des courbes et des surfaces. La formulation de leurs définitions générales n'entre pas dans le cadre de cet ouvrage. Le lecteur habitué aux définitions strictes peut entendre par courbe et surface respectivement une courbe algébrique et une surface algébrique, toutefois tous les résultats s'appliquent également au cas plus général.

3. Equations paramétriques des courbes. Supposons que la courbe est la trajectoire d'un point qui se déplace. A chaque instant t nous connaissons la position du point, autrement dit, ses coordonnées par rapport à un repère préalablement choisi. Cela signifie que les coordonnées (x, y, z) (ou (x, y) pour une courbe plane) sont des fonctions données du temps :

$$x = \varphi(t), \quad y = \psi(t), \quad z = \chi(t). \qquad (4)$$

Le fait que la variable t a ici le sens physique de temps n'est pas essentiel. Si l'on définit les coordonnées du point comme une fonction d'une variable quelconque ou, comme on dit, d'un *paramètre*, on définit par là même une courbe.

Les équations de la forme (4) sont appelées *équations paramétriques d'une courbe dans l'espace*, tandis que les équations de la forme $x = \varphi(t)$, $y = \psi(t)$, *équations paramétriques d'une courbe dans le plan*. Rappelons que l'équation de la courbe est une assertion sur les coordonnées des points qui n'est vraie que pour les points de la courbe. Dans le cas considéré, l'énoncé complet de cette assertion est le suivant : il existe un nombre t qui vérifie les égalités (4).

Par exemple, les équations $x = r \cos t$, $y = r \sin t$ définissent le cercle de rayon r et de centre à l'origine des coordonnées. On s'en assure aisément si l'on remarque que t est l'angle entre le rayon vecteur d'un point du cercle et le vecteur de base $\mathbf{e}_1$.

Les équations $x = r \cos t$, $y = r \sin t$, $z = at$ définissent une *hélice* située sur le cylindre de rayon r, dont le pas vaut $2\pi a$.

4. Equations paramétriques des surfaces. Cônes. Par analogie avec les équations paramétriques d'une courbe introduisons les équations paramétriques d'une surface. Les équations de la forme

$$x = \varphi(u, v), \quad y = \psi(u, v), \quad z = \chi(u, v) \tag{5}$$

sont appelées *équations paramétriques d'une surface* si à chaque point M de la surface correspond un couple de nombres (u, v) pour lequel les coordonnées de M s'obtiennent à partir de ces équations, et inversement, pour les points ne se trouvant pas sur la surface il n'existe pas de tel couple de nombres.

Cherchons à titre d'exemple les équations paramétriques d'un cône. On appelle *surface conique* ou *cône* une surface engendrée par des droites passant par un point fixe appelé *sommet du cône*. Les droites sont dénommées *génératrices* (fig. 15), quant à la courbe de la surface ne passant pas par le sommet et rencontrant toutes les génératrices, elle porte le nom de *directrice du cône*.

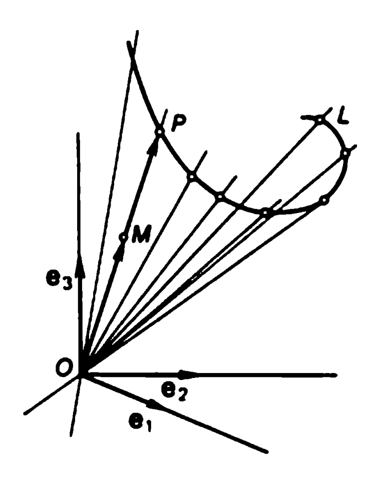


Fig. 15. L — directrice, MP — génératrice

Soient donnés le repère cartésien dont l'origine coïncide avec le sommet O du cône et les équations paramétriques de la directrice $x = f(u)$, $y = g(u)$, $z = h(u)$. Passant par un point arbitraire $M(x, y, z)$ du cône, la génératrice coupe la directrice en un point P de coordonnées $f(u)$, $g(u)$, $h(u)$, où u est la valeur correspondante du paramètre. Les vecteurs $\overrightarrow{OM}$ et $\overrightarrow{OP}$ étant colinéaires, il existe un nombre v tel que $\overrightarrow{OM} = v\overrightarrow{OP}$. Ecrivant cette égalité en coordonnées, on exprime les coordonnées d'un point arbitraire du cône en fonction des paramètres u et v :

$$x = v\,f(u),$$
$$y = v\,g(u),$$
$$z = v\,h(u).$$

On laisse au lecteur le soin de vérifier qu'aucun point extérieur au cône ne vérifie ces équations, quels que soient u et v.

5. Equations où l'une des coordonnées est absente. Considérons un cas particulier quand l'équation de la surface ne contient pas l'une des coordonnées, par exemple z, et par suite, prend la forme $G(x, y) = 0$. Supposons que le point M_0 de coordonnées x_0, y_0, z_0 se trouve sur la surface. Dans ce cas tout point de coordonnées x_0, y_0, z (où z est quelconque) appartient également à la surface. On voit sans peine que ces points engendrent une droite passant par le point M_0 en direction du vecteur $\mathbf{e}_3$. Donc, à tout point M_0 correspond sur la surface une droite passant par M_0 en direction de $\mathbf{e}_3$.

DÉFINITION. La surface constituée des droites parallèles à une direction donnée est appelée *surface cylindrique* ou *cylindre* et les droites, ses *génératrices* (fig. 16). La courbe de la surface rencontrant toutes les génératrices est dénommée *directrice.*

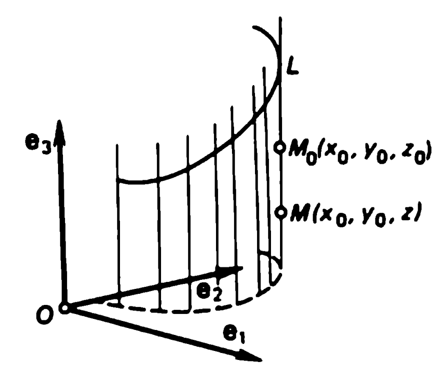


Fig. 16. L — directrice, M_0M — génératrice

On a montré que l'équation, dans laquelle l'une des coordonnées est absente, définit un cylindre dont les génératrices sont parallèles à l'axe des coordonnées correspondant.

A titre d'exemple on recommande au lecteur de construire une surface définie dans un repère cartésien rectangulaire de l'espace par l'équation $x^2 + y^2 = r^2$. Cette surface est un *cylindre circulaire droit.*

§ 2. Equations de la droite et du plan

Ce paragraphe ainsi que le suivant sont consacrés aux équations du plan, de la droite dans le plan et dans l'espace. Ces équations ont beaucoup de commun et on les étudiera ensemble. Au cas où les propositions semblables admettent les mêmes démonstrations on n'en donnera qu'une seule.

1. Surfaces et courbes du premier ordre. L'équation du premier degré, ou équation linéaire, liant les coordonnées d'un point dans l'espace est de la forme

$$Ax + By + Cz + D = 0, \tag{1}$$

où tous les coefficients ne sont pas nuls en même temps, c'est-à-dire $A^2 + B^2 + C^2 \neq 0$. De façon analogue, l'équation du premier degré, ou équation linéaire, liant les coordonnées d'un point dans le plan est l'équation

$$Ax + By + C = 0, \tag{2}$$

à la condition que $A^2 + B^2 \neq 0$.

On démontre que les surfaces et les courbes du premier ordre sont des plans et des droites. Plus précisément, on a les théorèmes suivants.

THÉORÈME 1. *Etant donné un repère cartésien de l'espace, à tout plan on peut associer une équation linéaire, et inversement, toute équation linéaire* (1) *définit un plan.*

THÉORÈME 2. *Etant donné un repère cartésien du plan, à toute droite on peut associer une équation linéaire, et inversement, toute équation linéaire* (2) *définit une droite.*

Les deux théorèmes se démontrent de façon analogue. Démontrons, par exemple, le théorème 1.

Soit donné un plan. Choisissons un repère dont l'origine et les deux premiers vecteurs de base $\mathbf{e}_1$ et $\mathbf{e}_2$ sont situés dans le plan donné, tandis que le vecteur $\mathbf{e}_3$ est arbitraire. Rapporté à ce repère, le plan a pour équation $z = 0$ qui est une équation linéaire. En vertu du théorème d'invariance de l'ordre, ce plan est aussi défini par une équation linéaire dans tout autre repère cartésien.

Inversement, soient donnés un repère cartésien quelconque et l'équation (1). Cherchons l'ensemble des points dont les coordonnées vérifient cette équation. Supposons pour fixer les idées que dans l'équation le coeffi-

cient C n'est pas nul et procédons au changement de repère en posant

$$x' = x, \quad y' = y, \quad z' = Ax + By + Cz + D. \tag{3}$$

Défini dans l'ancien repère par l'équation (1), l'ensemble cherché a pour équation $z' = 0$ dans le nouveau repère et, par suite, est un plan. Il nous reste à démontrer que les équations (3) définissent en effet le passage à un nouveau repère. Résolvons-les par rapport à x, y et z. Il vient

$$x = x', \quad y = y', \quad z = -\frac{A}{C}x' - \frac{B}{C}y' + \frac{z'}{C} - \frac{D}{C}.$$

En s'appuyant sur les formules (5) du § 4, ch. I, on peut affirmer que ces égalités définissent le passage du repère initial au repère $\{O', \mathbf{e}_1', \mathbf{e}_2', \mathbf{e}_3'\}$, avec $O'(0, 0, -D/C)$, $\mathbf{e}_1'(1, 0, -A/C)$, $\mathbf{e}_2'(0, 1, -B/C)$, $\mathbf{e}_3'(0, 0, 1/C)$. Les composantes des vecteurs $\mathbf{e}_1'$, $\mathbf{e}_2'$ et $\mathbf{e}_3'$ montrent que ces vecteurs ne sont pas coplanaires, quelles que soient les valeurs de A, B et C. La démonstration est ainsi achevée.

Les théorèmes 1 et 2 donnent une réponse exhaustive au problème des équations du plan et de la droite dans le plan. Mais, vu l'importance de ces équations, on en cherchera d'autres formes.

2. Equations paramétriques de la droite et du plan. Une droite (dans le plan ou dans l'espace) est complètement définie par l'un de ses points et par un vecteur non nul parallèle à cette droite. Il va de soi que le choix du point et du vecteur peut être effectué de nombreuses façons, mais en attendant on admettra qu'ils ont été choisis d'une certaine manière et on les appellera *point initial* et *vecteur directeur de la droite*. De façon analogue, un plan est défini par un point et deux vecteurs non colinéaires situés dans ce plan, appelés *point initial* et *vecteurs directeurs du plan*.

On admettra qu'est donné un repère cartésien dans l'espace ou dans le plan (si l'on étudie la droite en géométrie plane). Cela signifie en particulier qu'à chaque point on fait correspondre son rayon vecteur issu de l'origine des coordonnées.

Soit une droite. Notons $\mathbf{r}_0$ et $\mathbf{a}$ respectivement le rayon vecteur de son point initial M_0 et le vecteur directeur de la droite. Considérons maintenant un point quelconque M et désignons par $\mathbf{r}$ son rayon vecteur (fig. 17). Le vecteur $\mathbf{r} - \mathbf{r}_0$ dont l'origine M_0 se trouve sur la droite est parallèle à la droite si et seulement si son extrémité, le point M, est aussi située sur la droite. Dans ce cas et seulement dans ce cas, il existe pour le point M un nombre t tel que

$$\mathbf{r} - \mathbf{r}_0 = t\mathbf{a}. \tag{4}$$

Inversement, quel que soit le nombre t dans la formule (4), le vecteur $\mathbf{r}$ est le rayon vecteur d'un point qui appartient à la droite

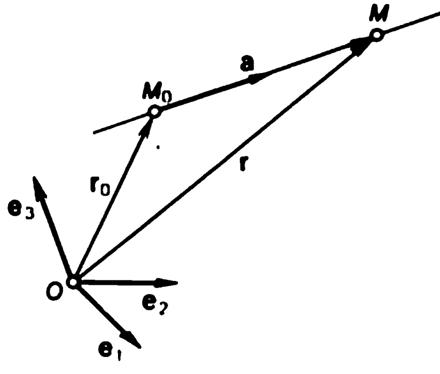


Fig. 17.

Dans la formule (4), la variable t parcourant toutes les valeurs réelles est appelée *paramètre*, et l'équation (4) *équation paramétrique vectorielle de la droite*.

L'équation paramétrique vectorielle de la droite est de la même forme en géométrie plane et dans l'espace mais, décomposée suivant les vecteurs de base, elle se réduit dans le premier cas à deux et dans le second cas à trois équations scalaires. En effet, désignons les coordonnées des points M et M_0 par (x, y) et (x_0, y_0) si l'on étudie une droite dans le plan et par (x, y, z) et (x_0, y_0, z_0) si la droite est dans l'espace. Notons respectivement (a_1, a_2) et (a_1, a_2, a_3) les composantes du vecteur $\mathbf{a}$ et décomposons les deux membres de l'égalité (4) suivant les vecteurs de base. Il vient

$$\left.\begin{aligned} x - x_0 &= a_1 t, \\ y - y_0 &= a_2 t \end{aligned}\right\} \tag{5}$$

ou

$$\left.\begin{aligned} x - x_0 &= a_1 t, \\ y - y_0 &= a_2 t, \\ z - z_0 &= a_3 t \end{aligned}\right\} \tag{6}$$

suivant le nombre de vecteurs composant la base. Les équations (5) sont appelées *équations paramétriques de la droite dans le plan* et les équations (6), *équations paramétriques de la droite dans l'espace*.

Cherchons maintenant les équations paramétriques du plan. Notons $\mathbf{r}_0$, $\mathbf{p}$ et $\mathbf{q}$ le rayon vecteur du point initial M_0 et les vecteurs directeurs du plan. Soit dans l'espace un point arbitraire M de rayon vecteur $\mathbf{r}$. L'origine du vecteur $\mathbf{r} - \mathbf{r}_0$ est dans le plan. Donc, son extrémité M est dans le plan si et seulement si ce vecteur se trouve dans le plan considéré. Les vecteurs $\mathbf{p}$ et $\mathbf{q}$ ne sont pas colinéaires (fig. 18). Par conséquent, si (et seulement si) le point M est dans le plan, il existe des nombres t_1 et t_2 tels que

$$\mathbf{r} - \mathbf{r}_0 = t_1 \mathbf{p} + t_2 \mathbf{q}. \tag{7}$$

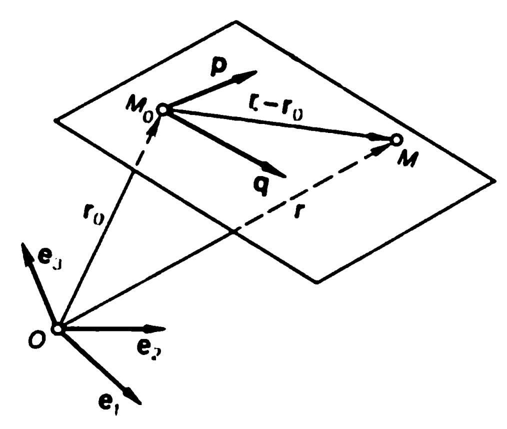


Fig. 18.

Cette équation porte le nom d'*équation paramétrique vectorielle du plan.* A tout point du plan correspondent deux valeurs des paramètres t_1 et t_2. Inversement, quelles que soient les valeurs des paramètres t_1 et t_2, l'équation (7) définit le rayon vecteur d'un point dans le plan.

Si (x, y, z) et (x_0, y_0, z_0) sont les coordonnées des points M et M_0, et (p_1, p_2, p_3) et (q_1, q_2, q_3) les composantes des vecteurs $\mathbf{p}$ et $\mathbf{q}$, en décomposant les deux membres de l'égalité (7) suivant les vecteurs de base, on obtient

$$\left.\begin{aligned} x - x_0 &= t_1 p_1 + t_2 q_1, \\ y - y_0 &= t_1 p_2 + t_2 q_2, \\ z - z_0 &= t_1 p_3 + t_2 q_3. \end{aligned}\right\} \qquad (8)$$

Les équations (8) sont appelées *équations paramétriques du plan.*

Signalons que le point initial et le vecteur directeur de la droite constituent sur cette droite son propre repère cartésien. La valeur du paramètre t correspondant à un point quelconque de la droite est la coordonnée de ce point par rapport à ce repère.

Une remarque analogue peut être faite pour le plan. Son point initial et les vecteurs directeurs y constituent son propre repère. Les valeurs des paramètres t_1 et t_2 correspondant à un point quelconque du plan sont les coordonnées de ce point par rapport à ce repère.

Etudions le passage des équations linéaires générales du plan et de la droite dans le plan à leurs équations paramétriques. Le point initial s'obtient aisément : si dans l'équation (1) du plan le coefficient A, par exemple, est différent de zéro, posons $y_0 = z_0 = 0$ et cherchons x_0 à partir de l'équation. On obtient ainsi le point $M_0\,(-D/A, 0, 0)$ dont les coordonnées vérifient l'équation (1), et qu'on peut prendre pour point initial. Pour une droite du plan les coordonnées du point initial se définissent de façon analogue.

Montrons maintenant comment trouver les vecteurs directeurs. Suppo-

sons qu'une droite du plan est définie par son équation générale $Ax + By + C = 0$ dans un repère cartésien quelconque et que (x_0, y_0) sont les coordonnées d'un point M_0 de la droite. En retranchant l'égalité $Ax_0 + By_0 + C = 0$ de l'équation de la droite, on obtient

$$A(x - x_0) + B(y - y_0) = 0.$$

Considérons maintenant le vecteur de composantes $(x - x_0, y - y_0)$. Son origine M_0 se trouve sur la droite et, par suite, il est parallèle à la droite si et seulement si son extrémité, le point M de coordonnées (x, y), se trouve sur la droite, et si l'équation précédente est vérifiée. En désignant les composantes du vecteur $\overrightarrow{M_0M}$ par α_1 et α_2, on peut énoncer la proposition qui suit.

PROPOSITION 1. *Tout vecteur non nul dont les composantes α_1 et α_2 vérifient l'équation $A\alpha_1 + B\alpha_2 = 0$ peut être pris pour le vecteur directeur de la droite $Ax + Bx + C = 0$. En particulier, le vecteur de composantes $(-B, A)$ est un vecteur directeur.*

De façon analogue on démontre la proposition suivante :

PROPOSITION 2. *Tout couple de vecteurs non colinéaires dont les composantes vérifient l'équation $A\alpha_1 + B\alpha_2 + C\alpha_3 = 0$ peut être pris pour les vecteurs directeurs du plan dont l'équation générale dans un repère cartésien est $Ax + By + Cz + D = 0$.*

3. Elimination d'un paramètre entre les équations paramétriques de la droite. Considérons dans le plan une droite définie par les équations paramétriques (5). Vu que le vecteur directeur n'est pas nul, l'une au moins de ses composantes a_1, a_2 est différente de zéro.

Supposons d'abord que $a_1 \neq 0$. Dans ce cas, on obtient de la première équation que $t = (x - x_0)/a_1$. Portant t dans la seconde équation on a

$$y - y_0 = \frac{a_2}{a_1}(x - x_0),$$

ou

$$y = kx + b, \tag{9}$$

avec $k = a_2/a_1$ et $b = y_0 - a_2x_0/a_1$. L'équation (9) est l'*équation de la droite résolue en ordonnée*. On peut aussi l'obtenir en résolvant l'équation $Ax + By + C = 0$ en y.

DÉFINITION. On appelle *coefficient angulaire de la droite* le rapport a_2/a_1 des composantes du vecteur directeur.

Posant $x = 0$ dans l'équation (9), on obtient $y = b$. Cela signifie que le point de coordonnées $(0, b)$ se trouve sur la droite (fig. 19). C'est le point d'intersection de la droite avec l'axe des ordonnées. Tout ce qui vient d'être dit peut être résumé dans la

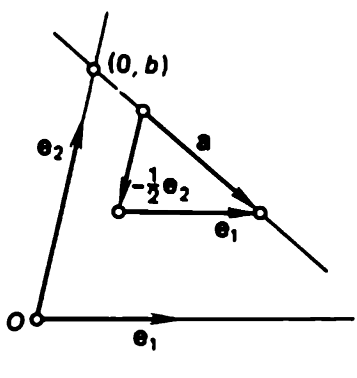


Fig. 19.

PROPOSITION 3. *Si la droite n'est pas parallèle à l'axe des ordonnées* ($a_1 \neq 0$), *son équation peut être écrite sous la forme* (9), *où k est le coefficient angulaire, et b l'ordonnée du point d'intersection de la droite avec l'axe des ordonnées.*

Considérons une équation résolue en ordonnée dans le repère cartésien rectangulaire. Dans cette équation, le terme b est égal en valeur absolue à la distance de l'origine des coordonnées au point d'intersection de la droite avec l'axe des ordonnées ; b est strictement positif si ce point se trouve du même côté de l'origine des coordonnées que l'extrémité du vecteur $\mathbf{e}_2$; dans le cas contraire, b est strictement négatif.

Le coefficient angulaire de la droite rapportée à un repère cartésien rectangulaire est égal à la tangente de l'angle que forme cette droite avec l'axe des abscisses. Cet angle est mesuré à partir de l'axe des abscisses dans le sens de la plus petite rotation amenant l'axe des abscisses sur celui des ordonnées (fig. 20).

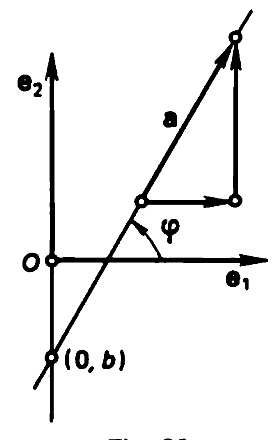


Fig. 20.

PROPOSITION 4. *Si la droite est parallèle à l'axe des ordonnées* ($a_1 = 0$), *son équation est de la forme* $x = x_0$, *où* x_0 *est l'abscisse du point d'intersection de la droite avec l'axe des abscisses.*

Pour le démontrer, il suffit de porter $a_1 = 0$ dans les équations (5). La première de ces équations prend alors la forme $x = x_0$, et la seconde se réduit à l'assertion que y est arbitraire. Autrement dit, la droite est constituée des points de coordonnées (x_0, y) où y parcourt toutes les valeurs possibles.

Chassons maintenant le paramètre t des équations paramétriques (6) de la droite dans l'espace. Posons d'abord qu'aucune des composantes du vecteur directeur n'est nulle. Alors

$$t = \frac{x - x_0}{a_1}, \quad t = \frac{y - y_0}{a_2}, \quad t = \frac{z - z_0}{a_3},$$

d'où l'on obtient deux égalités

$$\frac{x - x_0}{a_1} = \frac{y - y_0}{a_2} = \frac{z - z_0}{a_3}, \tag{10}$$

que vérifient les coordonnées de tout point de la droite et ne vérifient les coordonnées d'aucun point non situé sur la droite. Vu que toute droite dans l'espace peut être représentée, en accord avec la proposition 1 du § 1, comme une droite d'intersection de deux plans, elle se définit bien par un système de deux équations linéaires.

Si l'une des composantes du vecteur directeur, par exemple a_1, devient nulle, l'équation de la droite prend la forme

$$x = x_0, \quad \frac{y - y_0}{a_2} = \frac{z - z_0}{a_3}. \tag{10'}$$

Cette droite se trouve dans le plan $x = x_0$ et, par suite, est parallèle au plan $x = 0$. Le lecteur écrira aisément l'équation de la droite si s'annulle non pas a_1 mais une autre composante.

Lorsque deux composantes du vecteur directeur sont nulles, par exemple a_1 et a_2, la droite a pour équation

$$x = x_0, \quad y = y_0. \tag{10''}$$

Cette droite est parallèle à l'un des axes de coordonnées, dans notre cas à l'axe des cotes.

On représente souvent l'équation d'une droite sous la forme (10) en posant le numérateur égal à zéro si le dénominateur s'annule.

4. Equations vectorielles du plan et de la droite. Le plan est entièrement défini par son point initial et par un vecteur non nul $\mathbf{n}$ perpendiculaire au plan, appelé *vecteur normal*. Si $\mathbf{r}$ est le rayon vecteur d'un point du plan, $\mathbf{r} - \mathbf{r}_0$ se trouve dans le plan et, par suite,

$$(\mathbf{r} - \mathbf{r}_0, \mathbf{n}) = 0. \tag{11}$$

Inversement, on voit aisément que le point de rayon vecteur **r** appartient au plan si **r** vérifie (11). L'égalité (11) sera appelée *équation vectorielle du plan*.

Si **p** et **q** sont des vecteurs directeurs du plan, le vecteur $[\mathbf{p}, \mathbf{q}]$ est un vecteur normal. On aboutit ainsi à l'équation vectorielle de la forme

$$(\mathbf{r} - \mathbf{r}_0, \mathbf{p}, \mathbf{q}) = 0. \qquad (11')$$

On peut aussi donner à l'équation vectorielle du plan la forme

$$(\mathbf{r}, \mathbf{n}) + D = 0, \qquad (12)$$

où $D = -(\mathbf{r}_0, \mathbf{n})$. Cette équation ne renferme pas de rayon vecteur du point initial.

PROPOSITION 5. *Soient x, y, z les composantes du vecteur* **r** *par rapport à un repère cartésien donné. Le produit scalaire* $(\mathbf{r} - \mathbf{r}_0, \mathbf{n})$, *où* $\mathbf{n} \neq \mathbf{0}$, *s'écrit alors sous la forme du polynôme linéaire* $Ax + By + Cz + D$, *avec* $A^2 + B^2 + C^2 \neq 0$.

Inversement, étant donné un repère cartésien, il existe pour tout polynôme linéaire des vecteurs $\mathbf{r}_0$ *et* $\mathbf{n} \neq \mathbf{0}$ *tels que* $Ax + By + Cz + D = (\mathbf{r} - \mathbf{r}_0, \mathbf{n})$.

La première partie de la proposition est évidente. En effet, portons dans l'expression donnée le vecteur **r** décomposé suivant les vecteurs de base :

$$(x\mathbf{e}_1 + y\mathbf{e}_2 + z\mathbf{e}_3 - \mathbf{r}_0, \mathbf{n})$$

et chassons les parenthèses. On obtient le polynôme $Ax + By + Cz + D$ dans lequel $D = -(\mathbf{r}_0, \mathbf{n})$ et

$$A = (\mathbf{e}_1, \mathbf{n}), \quad B = (\mathbf{e}_2, \mathbf{n}), \quad C = (\mathbf{e}_3, \mathbf{n}). \qquad (13)$$

A, B, C ne sont pas simultanément nuls, car le vecteur non nul **n** ne peut être orthogonal à la fois à tous les vecteurs de base.

Pour démontrer l'assertion inverse, proposons-nous d'abord de trouver le vecteur **n** à partir des égalités (13) en supposant A, B et C donnés. Cherchons ce vecteur sous la forme

$$\mathbf{n} = \alpha[\mathbf{e}_2, \mathbf{e}_3] + \beta[\mathbf{e}_3, \mathbf{e}_1] + \gamma[\mathbf{e}_1, \mathbf{e}_2].$$

En multipliant scalairement cette égalité par $\mathbf{e}_1$ on obtient $(\mathbf{e}_1, \mathbf{n}) = \alpha(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. D'où $\alpha = A/(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$. De façon analogue, on obtient β et γ. Le vecteur **n** n'est pas nul, car $[\mathbf{e}_2, \mathbf{e}_3]$, $[\mathbf{e}_3, \mathbf{e}_1]$ et $[\mathbf{e}_1, \mathbf{e}_2]$ sont linéairement indépendants (proposition 7, § 3, ch. I) et α, β et γ ne sont pas simultanément nuls, comme d'ailleurs A, B et C.

Ainsi donc, on peut écrire le polynôme donné sous la forme

$$x(\mathbf{e}_1, \mathbf{n}) + y(\mathbf{e}_2, \mathbf{n}) + z(\mathbf{e}_3, \mathbf{n}) + D = (\mathbf{r}, \mathbf{n}) + D.$$

Le vecteur $\mathbf{r}_0$ doit satisfaire à la condition $-(\mathbf{r}_0, \mathbf{n}) = D$. Cherchons-le sous la forme $\mathbf{r}_0 = \lambda\mathbf{n}$. Après avoir fait la substitution on trouve $-\lambda(\mathbf{n}, \mathbf{n}) = D$, d'où $\mathbf{r}_0 = -D\mathbf{n}/|\mathbf{n}|^2$.

Remarquons que la proposition démontrée entraîne aisément le théorème 1. En outre, comme corollaire on obtient la proposition suivante :

PROPOSITION 6. *Si le repère cartésien est rectangulaire, le vecteur de composantes A, B et C est le vecteur normal du plan $Ax + By + Cz + D = 0$.*

Cela découle immédiatement des formules (13) en vertu de la proposition 1 du § 3, ch. I.

Dans le cas d'un repère cartésien quelconque, A, B, C sont les composantes du vecteur $\mathbf{n}$ suivant les vecteurs $\mathbf{e}_1^*$, $\mathbf{e}_2^*$, $\mathbf{e}_3^*$ de la base biorthogonale à $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

Tout ce qui a été dit plus haut sur le plan peut presque sans changement être dit de la droite dans le plan. Si $\mathbf{r}_0$ est le rayon vecteur du point initial de la droite et $\mathbf{n}$ un vecteur non nul perpendiculaire à cette droite, l'équation de la droite dans le plan s'écrit sous la forme

$$(\mathbf{r} - \mathbf{r}_0, \mathbf{n}) = 0 \quad \text{ou} \quad (\mathbf{r}, \mathbf{n}) + C = 0.$$

En coordonnées, cette équation est de la forme $Ax + By + C = 0$, où

$$A = (\mathbf{e}_1, \mathbf{n}), \quad B = (\mathbf{e}_2, \mathbf{n}).$$

PROPOSITION 7. *Si le repère cartésien est rectangulaire, le vecteur de composantes (A, B) est perpendiculaire à la droite d'équation $Ax + By + C = 0$.*

Dans cette proposition, comme plus haut dans la proposition 6, il est très important que le repère soit supposé cartésien rectangulaire. Pour souligner que cette condition est importante, il suffit de considérer la droite $x + y = 0$ et le vecteur de composantes (1, 1) qui sont représentés sur la figure 21 dans le cas où les vecteurs de base sont de longueur différente.

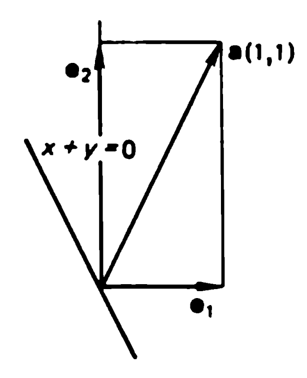


Fig. 21.

Passons maintenant à l'équation vectorielle de la droite dans l'espace. Elle peut être écrite sous la forme

$$[\mathbf{r} - \mathbf{r}_0, \mathbf{a}] = \mathbf{0}, \tag{14}$$

où $\mathbf{a}$ est le vecteur directeur de la droite et $\mathbf{r}_0$ le rayon vecteur du point initial de la droite. En effet, cette équation, comme l'équation paramétrique vectorielle de la droite, exprime la colinéarité des vecteurs $\mathbf{r} - \mathbf{r}_0$ et $\mathbf{a}$. Si l'on désigne $[\mathbf{r}_0, \mathbf{a}]$ par $\mathbf{d}$, on obtient l'équation vectorielle de la droite sans point initial :

$$[\mathbf{r}, \mathbf{a}] = \mathbf{d}. \tag{15}$$

Si la droite est définie par l'équation (15), on est en mesure de trouver le point initial sous la forme, par exemple, de $\mathbf{r}_0 = t[\mathbf{a}, \mathbf{d}]$. En portant $\mathbf{r}_0$ dans (15) on obtient $t[[\mathbf{a}, \mathbf{d}], \mathbf{a}] = \mathbf{d}$, d'où, en vertu de la formule du produit vectoriel double, $t\{(\mathbf{a}, \mathbf{a})\mathbf{d} - (\mathbf{a}, \mathbf{d})\mathbf{a}\} = \mathbf{d}$. Puisque $(\mathbf{a}, \mathbf{d}) = 0$, il s'ensuit que $t = (\mathbf{a}, \mathbf{a})^{-1}$ et $\mathbf{r}_0 = |\mathbf{a}|^{-2}[\mathbf{a}, \mathbf{d}]$.

Remarquons que le point $\mathbf{r}_0$ ainsi défini est le pied de la perpendiculaire abaissée de l'origine des coordonnées sur la droite, tandis que $\mathbf{r} = t[\mathbf{a}, \mathbf{d}]$ est l'équation paramétrique de cette perpendiculaire.

5. Conditions de parallélisme de deux plans et de deux droites dans un plan. La proposition 1 entraîne aisément la condition de parallélisme de deux droites dans un plan, imposée aux coefficients de leurs équations. La condition analogue pour les plans découle de la proposition 2 et des formules (13). Il faut toutefois avoir en vue que cette condition est satisfaite non seulement lorsque les équations définissent des droites parallèles différentes mais aussi dans le cas où les équations considérées définissent une même droite. Il est commode alors de dire que les droites définies par les équations se confondent.

PROPOSITION 8. 1) *Deux droites définies dans un repère cartésien par les équations*

$$Ax + By + C = 0, \quad A_1 x + B_1 y + C_1 = 0$$

sont parallèles si et seulement si les coefficients correspondants des variables x et y dans leurs équations sont proportionnels, c'est-à-dire s'il existe un nombre λ tel que

$$A_1 = \lambda A, \quad B_1 = \lambda B. \tag{16}$$

2) *Les droites se confondent si et seulement si leurs équations sont proportionnelles, c'est-à-dire si avec l'égalité* (16) *on a (pour le même λ) l'égalité*

$$C_1 = \lambda C. \tag{17}$$

DÉMONSTRATION. La première partie de la proposition découle immédiatement du fait que les vecteurs de composantes $(-B, A)$ et $(-B_1, A_1)$ sont des vecteurs directeurs des droites correspondantes.

Démontrons la seconde partie. Dans les égalités (16) et (17), λ ne peut être nul, vu que les coefficients de l'équation de la droite ne peuvent s'annuler simultanément. Par conséquent, si ces égalités sont vérifiées, les équations sont équivalentes et définissent une même droite.

Vérifions la proposition réciproque. Si les droites sont parallèles, leurs équations, comme on l'a vu, doivent avoir la forme $Ax + By + C = 0$ et $\lambda(Ax + By) + C_1 = 0$ pour un certain λ. Si de plus un point de coordonnées (x_0, y_0) appartient à deux droites, les égalités $Ax_0 + By_0 + C = 0$ et $\lambda(Ax_0 + By_0) + C_1 = 0$ sont aussi vérifiées. De ces égalités il résulte que $C_1 = \lambda C$. La proposition est démontrée.

PROPOSITION 9. *Soient P et Q deux plans définis dans un repère cartésien par les équations*

$$Ax + By + Cz + D = 0, \quad A_1 x + B_1 y + C_1 z + D = 0.$$

Pour qu'ils soient parallèles il faut et il suffit qu'il existe un nombre λ tel que

$$A_1 = \lambda A, \quad B_1 = \lambda B, \quad C_1 = \lambda C. \tag{18}$$

Les plans P et Q se confondent si et seulement si, outre la relation (18), *est vérifiée* (*pour le même* λ) *l'égalité*

$$D_1 = \lambda D. \tag{19}$$

DÉMONSTRATION. La condition est *nécessaire*. En effet, selon la proposition 5, les équations des plans P et Q peuvent être représentées sous la forme $(\mathbf{r} - \mathbf{r}_0, \mathbf{n}) = 0$ et $(\mathbf{r} - \mathbf{r}_1, \mathbf{n}_1) = 0$, où $\mathbf{n}$ et $\mathbf{n}_1$ sont des vecteurs liés aux coefficients des équations par les formules (13). Si les plans sont parallèles, $\mathbf{n}$ et $\mathbf{n}_1$ sont colinéaires et, par suite, il existe un nombre λ tel que $\mathbf{n}_1 = \lambda\mathbf{n}$. En vertu de (13) on a $A_1 = (\mathbf{e}_1, \mathbf{n}_1) = \lambda(\mathbf{e}_1, \mathbf{n}) = \lambda A$. On obtient de façon analogue les autres égalités de (18). Posons maintenant que P et Q coïncident. Alors d'après ce qui a été démontré plus haut, l'équation de Q est de la forme $\lambda(Ax + By + Cz) + D_1 = 0$. Vu que les plans possèdent un point commun, il découle de

$$\lambda(Ax_0 + By_0 + Cz_0) + D_1 = 0 \quad \text{et} \quad Ax_0 + By_0 + Cz_0 + D = 0$$

que $D_1 = \lambda D$.

La condition est *suffisante*. Il nous faut démontrer que deux équations satisfaisant aux conditions (18) et (19) sont des équations d'un même plan et que les équations satisfaisant à la condition (18) définissent des plans parallèles. La première assertion est évidente, vu que ces équations sont

équivalentes. La seconde assertion résulte de la proposition 2. En effet, si les conditions (18) sont satisfaites, les composantes des vecteurs directeurs des deux plans doivent vérifier la même équation. Or cela signifie que le même couple de vecteurs joue le rôle de vecteurs directeurs de l'un et de l'autre plan. La proposition est démontrée.

Les conditions (16) ne signifient rien d'autre que le fait que les vecteurs de composantes (A, B) et (A_1, B_1) sont colinéaires. De façon analogue, les conditions (18) affirment la colinéarité des vecteurs de composantes (A, B, C) et (A_1, B_1, C_1). D'où la possibilité d'écrire la condition de parallélisme de deux droites du plan sous la forme

$$\begin{vmatrix} A & B \\ A_1 & B_1 \end{vmatrix} = 0, \tag{16'}$$

et la condition de parallélisme de deux plans sous la forme

$$\begin{vmatrix} B & C \\ B_1 & C_1 \end{vmatrix} = \begin{vmatrix} C & A \\ C_1 & A_1 \end{vmatrix} = \begin{vmatrix} A & B \\ A_1 & B_1 \end{vmatrix} = 0. \tag{18'}$$

(Voir les propositions 9 et 10, § 3, ch. I.)

La proposition 8 peut être énoncée sous une forme purement algébrique si l'on tient compte du fait que les coordonnées du point d'intersection des droites représentent la solution du système de leurs équations.

PROPOSITION 10. *Si*

$$\begin{vmatrix} A & B \\ A_1 & B_1 \end{vmatrix} = 0,$$

le système d'équations

$$Ax + By + C = 0, \quad A_1 x + B_1 y + C_1 = 0$$

n'a pas de solutions ou en possède une infinité (suivant C et C_1). Si

$$\begin{vmatrix} A & B \\ A_1 & B_1 \end{vmatrix} \neq 0,$$

le système a une solution unique (x, y) quels que soient C et C_1.

Il va de soi que la proposition 10 se démontre aisément de façon directe sans recourir à des considérations géométriques. Pour un cas plus général on effectuera cette démonstration au chapitre V. Une telle démonstration n'est qu'une autre démonstration de la proposition 8.

6. Equations de la droite dans l'espace. Une droite dans l'espace peut être définie comme l'intersection de deux plans et, par suite, dans un repère cartésien quelconque se définit par le système d'équations

$$Ax + By + Cz + D = 0, \quad A_1 x + B_1 y + C_1 z + D_1 = 0. \tag{20}$$

La condition de parallélisme de deux plans permet d'indiquer quand le système (20) définit une droite unique. L'intersection de deux plans est une droite si et seulement s'ils ne sont pas parallèles (et, en particulier, ne se confondent pas). Aussi le système (20) définit-il une droite unique si et seulement si l'un au moins des déterminants (18′) est différent de zéro, c'est-à-dire si

$$\begin{vmatrix} B & C \\ B_1 & C_1 \end{vmatrix}^2 + \begin{vmatrix} C & A \\ C_1 & A_1 \end{vmatrix}^2 + \begin{vmatrix} A & B \\ A_1 & B_1 \end{vmatrix}^2 \neq 0. \qquad (21)$$

Il va de soi que le système (20) peut être remplacé par tout système équivalent. Dans ce cas la droite sera représentée par l'intersection de deux autres plans passant par elle. Les équations (10) la représentent comme l'intersection des plans

$$\frac{y - y_0}{a_2} = \frac{z - z_0}{a_3}, \quad \frac{x - x_0}{a_1} = \frac{y - y_0}{a_2}$$

respectivement parallèles aux axes d'abscisses et de cotes.

Il est important de savoir trouver un point initial et un vecteur directeur de la droite définie par le système d'équations linéaires. La condition (21) signifie que des trois déterminants d'ordre 2 un au moins est différent de zéro. Posons pour fixer les idées que $AB_1 - A_1B \neq 0$. Pour obtenir un point initial de la droite, considérons le système (20) comme un système d'équations en x et y. Donnons à la variable z une valeur numérique arbitraire, par exemple 0. On a alors pour x et y le système d'équations

$$Ax + By + D = 0, \quad A_1x + B_1y + D_1 = 0.$$

En vertu de la proposition 10, ce système a une solution qu'on notera (x_0, y_0). Le point M_0 de coordonnées $(x_0, y_0, 0)$ se trouve alors sur la droite car ses coordonnées vérifient le système (20).

Cherchons maintenant le vecteur directeur. Si le repère cartésien est rectangulaire, les vecteurs de composantes A, B, C et A_1, B_1, C_1 sont perpendiculaires aux plans $Ax + By + Cz + D = 0$ et $A_1x + B_1y + C_1z + D_1 = 0$. Donc, le produit vectoriel de ces vecteurs est parallèle à la droite (20) suivant laquelle les plans se coupent. En calculant les composantes du produit vectoriel dans une base orthonormée, on obtient les composantes suivantes du vecteur directeur de la droite :

$$\begin{vmatrix} B & C \\ B_1 & C_1 \end{vmatrix}, \quad \begin{vmatrix} C & A \\ C_1 & A_1 \end{vmatrix}, \quad \begin{vmatrix} A & B \\ A_1 & B_1 \end{vmatrix}. \qquad (22)$$

PROPOSITION 11. *Le vecteur de composantes* (22) *est le vecteur directeur de la droite* (20) *quel que soit le repère cartésien.*

DÉMONSTRATION. Selon la proposition 2, tout vecteur non nul dont les composantes $\alpha_1, \alpha_2, \alpha_3$ vérifient l'équation $A\alpha_1 + B\alpha_2 + C\alpha_3 = 0$ est parallèle au plan $Ax + By + Cz + D = 0$. Si en outre $\alpha_1, \alpha_2, \alpha_3$ vérifient l'équation $A_1\alpha_1 + B_1\alpha_2 + C_1\alpha_3 = 0$, le vecteur est aussi parallèle au plan $A_1x + B_1y + C_1z + D_1 = 0$ et, partant, peut être pris pour vecteur directeur de la droite (20). Le vecteur de composantes (22) n'est pas nul en vertu de la relation (21). Il est aisé de vérifier immédiatement que ses composantes satisfont aux deux conditions posées plus haut. La démonstration est ainsi achevée.

§ 3. Quelques problèmes sur les droites et les plans

1. Equation d'une droite passant par deux points. Soient donnés dans l'espace un repère cartésien quelconque et deux points M_1 et M_2 de coordonnées (x_1, y_1, z_1) et (x_2, y_2, z_2). Pour écrire l'équation de la droite passant par M_1 et M_2 il suffit de prendre M_1 pour son point initial et $\overrightarrow{M_1M_2}$ pour son vecteur directeur. Ce vecteur n'est pas nul si les points ne se confondent pas. D'après la formule (10) du § 2 il vient

$$\frac{x - x_1}{x_2 - x_1} = \frac{y - y_1}{y_2 - y_1} = \frac{z - z_1}{z_2 - z_1}. \tag{1}$$

Si dans ces égalités l'un quelconque des dénominateurs est nul, on doit égaler à zéro le numérateur correspondant.

En géométrie plane le problème est résolu de la même façon. La différence réside dans le fait que les coordonnées des points sont maintenant (x_1, y_1) et (x_2, y_2), et l'on obtient (selon les propositions 3 et 4 du § 2) l'équation

$$y - y_1 = \frac{y_2 - y_1}{x_2 - x_1}(x - x_1) \tag{2}$$

si $x_1 \neq x_2$, et

$$x = x_1$$

si $x_1 = x_2$.

2. Equation d'un plan passant par trois points. Soient dans un repère cartésien trois points non alignés M_1, M_2, M_3 de coordonnées (x_1, y_1, z_1), (x_2, y_2, z_2) et (x_3, y_3, z_3). Pour écrire l'équation du plan passant par ces points choisissons le point M_1 pour point initial et les vecteurs $\overrightarrow{M_1M_2}$ et $\overrightarrow{M_1M_3}$ pour vecteurs directeurs. Alors selon les formules (11′) du § 2 et les formules (10) du § 3, ch. I, il vient

$$\begin{vmatrix} x - x_1 & y - y_1 & z - z_1 \\ x_2 - x_1 & y_2 - y_1 & z_2 - z_1 \\ x_3 - x_1 & y_3 - y_1 & z_3 - z_1 \end{vmatrix} = 0.$$

3. Conditions de parallélisme d'une droite et d'un plan. Supposons que la droite est définie par l'équation $\mathbf{r} - \mathbf{r}_0 = t\mathbf{a}$, et le plan par l'une des équations $(\mathbf{r} - \mathbf{r}_1, \mathbf{n}) = 0$, $(\mathbf{r} - \mathbf{r}_1, \mathbf{p}, \mathbf{q}) = 0$ ou $\mathbf{r} - \mathbf{r}_1 = t_1\mathbf{p} + t_2\mathbf{q}$. La droite est parallèle au plan (ou peut-être s'y trouve) si

$$(\mathbf{a}, \mathbf{n}) = 0 \quad \text{ou} \quad (\mathbf{a}, \mathbf{p}, \mathbf{q}) = 0.$$

Posons que le plan est défini par l'équation linéaire $Ax + By + Cz + D = 0$. Il vient

$$(\mathbf{n}, \mathbf{a}) = (\mathbf{n}, \mathbf{e}_1)\, a_1 + (\mathbf{n}, \mathbf{e}_2)\, a_2 + (\mathbf{n}, \mathbf{e}_3)\, a_3 = 0.$$

D'où, en appliquant les formules (13) du § 2, on obtient

$$Aa_1 + Ba_2 + Ca_3 = 0. \tag{3}$$

Posons que la droite est définie par deux équations linéaires

$$A_1 x + B_1 y + C_1 z + D_1 = 0, \quad A_2 x + B_2 y + C_2 z + D_2 = 0.$$

Alors en se conformant à la proposition 11 du § 2, on peut poser

$$a_1 = \begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix}, \quad a_2 = \begin{vmatrix} C_1 & A_1 \\ C_2 & A_2 \end{vmatrix}, \quad a_3 = \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix},$$

et la condition (3) s'écrit sous la forme

$$A \begin{vmatrix} B_1 & C_1 \\ B_2 & C_2 \end{vmatrix} + B \begin{vmatrix} C_1 & A_2 \\ C_2 & A_2 \end{vmatrix} + C \begin{vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{vmatrix} = 0,$$

ou bien

$$\begin{vmatrix} A & B & C \\ A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \end{vmatrix} = 0. \tag{4}$$

On vérifie aisément que toutes les conditions mentionnées ici sont non seulement nécessaires mais aussi suffisantes.

Il découle de la formule (4) que trois plans se coupent en un point si et seulement si les coefficients de leurs équations satisfont à la condition

$$\begin{vmatrix} A & B & C \\ A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \end{vmatrix} \neq 0. \tag{5}$$

En effet, cette inégalité signifie que la droite d'intersection de deux quelconques de ces plans n'est pas parallèle au troisième plan.

4. Equations du plan et de la droite en fonction de leurs coordonnées à l'origine. On appelle *équation du plan en fonction de ses coordonnées à*

l'origine l'équation de la forme

$$\frac{x}{a} + \frac{y}{b} + \frac{z}{c} = 1. \tag{6}$$

PROPOSITION 1. *Si le plan est défini par l'équation* (6) *dans un repère cartésien rectangulaire, les nombres a, b, c sont égaux en valeur absolue à la longueur des segments interceptés par le plan sur les axes de coordonnées* (fig. 22). *Le signe de ces nombres dépend de la position du segment correspondant sur le demi-axe* (*positif ou négatif*).

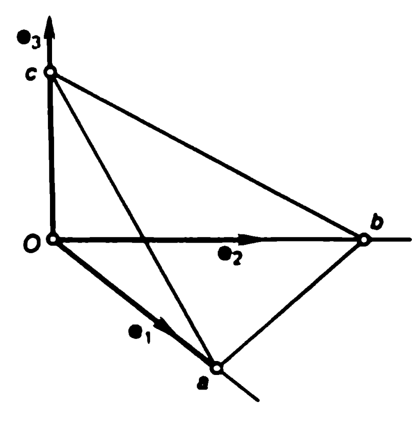


Fig. 22.

Pour justifier cette proposition il suffit de porter dans l'équation (6) les triplets de nombres suivants : $(a, 0, 0)$, $(0, b, 0)$ et $(0, 0, c)$.

L'équation de la droite dans le plan en fonction de ses coordonnées à l'origine est de la forme

$$\frac{x}{a} + \frac{y}{b} = 1. \tag{7}$$

Cette équation vérifie une proposition analogue à la proposition 1.

5. Demi-espace. Soient un plan P et le vecteur normal $\mathbf{n}$ de P. On appelle *demi-espace* défini par P et $\mathbf{n}$ l'ensemble des points M tels que, étant donné un point M_0 du plan P, l'angle des vecteurs $\mathbf{n}$ et $\overrightarrow{M_0M}$ soit au plus égal à $\pi/2$.

Si $\mathbf{r}$ est le rayon vecteur du point M, et $\mathbf{r}_0$ du point M_0, la définition du demi-espace est équivalente à l'inégalité $(\mathbf{r} - \mathbf{r}_0, \mathbf{n}) \geqslant 0$. Cette inégalité est justement l'*équation du demi-espace*.

En utilisant ce fait, il est aisé de vérifier que la définition du demi-espace ne dépend pas du choix du point M_0 dans le plan. En effet, si $M_0'(\mathbf{r}_0')$ est un autre point du plan, le vecteur $\mathbf{a} = \mathbf{r}_0 - \mathbf{r}_0'$ est perpendiculaire à $\mathbf{n}$, et on a

$$(\mathbf{r} - \mathbf{r}_0', \mathbf{n}) = (\mathbf{r} - \mathbf{r}_0 + \mathbf{a}, \mathbf{n}) = (\mathbf{r} - \mathbf{r}_0, \mathbf{n}).$$

On obtient l'équation du demi-espace en coordonnées si l'on se souvient que l'expression $(\mathbf{r} - \mathbf{r}_0, \mathbf{n})$ s'écrit en coordonnées au moyen du polynôme $Ax + By + Cz + D$. Ainsi donc, dans le repère cartésien, le demi-espace est défini par l'inéquation linéaire

$$Ax + By + Cz + D \geqslant 0.$$

Inversement, selon la proposition 5 du § 2, toute inéquation linéaire peut être écrite sous la forme $(\mathbf{r} - \mathbf{r}_0, \mathbf{n}) \geqslant 0$, d'où l'on obtient immédiatement qu'elle définit un demi-espace.

Le plan P et le vecteur $\mathbf{n}' = -\mathbf{n}$ déterminent un autre demi-espace défini par l'inéquation $(\mathbf{r} - \mathbf{r}_0, \mathbf{n}') \geqslant 0$, ou $(\mathbf{r} - \mathbf{r}_0, \mathbf{n}) \leqslant 0$. Il peut être appelé demi-espace « négatif » pour le différencier du demi-espace « positif » $(\mathbf{r} - \mathbf{r}_0, \mathbf{n}) \geqslant 0$. Mais cette appellation est conventionnelle vu qu'elle dépend du choix du vecteur $\mathbf{n}$. Le changement du sens de ce vecteur équivaut au changement des signes des coefficients et du terme constant dans l'équation du plan ou à une multiplication de l'équation du plan par un facteur négatif. Avec cela, le demi-espace « positif » devient « négatif » et réciproquement.

Si $M_1(x_1, y_1, z_1)$ et $M_2(x_2, y_2, z_2)$ sont deux points non situés dans le plan donné, et $Ax_1 + By_1 + Cz_1 + D$ et $Ax_2 + By_2 + Cz_2 + D$ les résultats de substitution de leurs coordonnées dans le premier membre de l'équation du plan, la coïncidence ou la non-coïncidence des signes de ces nombres est indépendante du choix du vecteur $\mathbf{n}$ ou de la multiplication de l'équation du plan par un facteur numérique. Dans le cas de coïncidence des signes, les points M_1 et M_2 se trouvent dans un même demi-espace.

Dans la résolution des problèmes, la remarque suivante peut présenter de l'intérêt : si le point de coordonnées x_0, y_0, z_0 est dans le plan, le point de coordonnées $x_0 + A$, $y_0 + B$ et $z_0 + C$ se trouve dans le demi-espace « positif ». Autrement dit, le vecteur de composantes A, B, C est toujours orienté vers le demi-espace « positif ». Le lecteur peut le vérifier aisément par substitution immédiate.

D'une façon analogue à ce qui vient d'être dit sur les demi-espaces, on peut définir un *demi-plan* et démontrer que l'inéquation $Ax + By + C \geqslant 0$, qui relie les coordonnées cartésiennes d'un point dans le plan, définit un demi-plan. Le second demi-plan ayant pour frontière la droite $Ax + By + C = 0$ est défini par l'inéquation $Ax + By + C \leqslant 0$.

6. Distance d'un point à un plan. Soient donnés un plan d'équation $(\mathbf{r} - \mathbf{r}_0, \mathbf{p}, \mathbf{q}) = 0$ et un point M de rayon vecteur $\mathbf{R}$. La distance du point M au plan se détermine le plus aisément si l'on divise le volume du parallélépipède construit sur les vecteurs $\mathbf{R} - \mathbf{r}_0$, $\mathbf{p}$ et $\mathbf{q}$ par l'aire de sa base

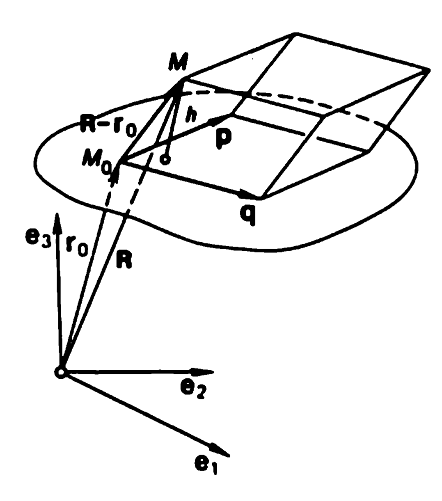


Fig. 23.

(fig. 23). On obtient

$$h = \frac{|(\mathbf{R} - \mathbf{r}_0, \mathbf{p}, \mathbf{q})|}{|[\mathbf{p}, \mathbf{q}]|} .$$

Pour tout vecteur $\mathbf{n}$ normal au plan on peut choisir les vecteurs directeurs $\mathbf{p}$ et $\mathbf{q}$ de façon que $[\mathbf{p}, \mathbf{q}] = \mathbf{n}$. Aussi pour tout vecteur normal $\mathbf{n}$ a-t-on

$$h = \frac{|(\mathbf{R} - \mathbf{r}_0, \mathbf{n})|}{|\mathbf{n}|} . \tag{8}$$

Si dans le repère cartésien rectangulaire le point M a pour coordonnées X, Y, Z, l'égalité (8) s'écrira (proposition 6, § 2)

$$h = \frac{|AX + BY + CZ + D|}{\sqrt{A^2 + B^2 + C^2}} . \tag{9}$$

Ici $D = -(\mathbf{r}_0, \mathbf{n})$ et A, B, C sont les composantes de $\mathbf{n}$.

Il est évident que $h = 0$ si et seulement si M est dans le plan. L'équation

$$\frac{Ax + By + Cx + D}{\sqrt{A^2 + B^2 + C^2}} = 0$$

s'appelle *équation normée du plan.*

PROPOSITION 2. *La distance d'un point à un plan est égale à la valeur absolue du résultat de substitution de ses coordonnées dans l'équation normée du plan.*

7. Distance d'un point à une droite. Si la droite est définie par l'équation $[\mathbf{r} - \mathbf{r}_0, \mathbf{a}] = \mathbf{0}$, on est en mesure de trouver la distance h du point de rayon vecteur $\mathbf{R}$ à cette droite en divisant l'aire du parallélogramme construit sur les vecteurs $\mathbf{R} - \mathbf{r}_0$ et $\mathbf{a}$ par la longueur de sa base (fig. 24). Le résultat peut être exprimé par la formule

$$h = \frac{|[\mathbf{R} - \mathbf{r}_0, \mathbf{a}]|}{|\mathbf{a}|} . \tag{10}$$

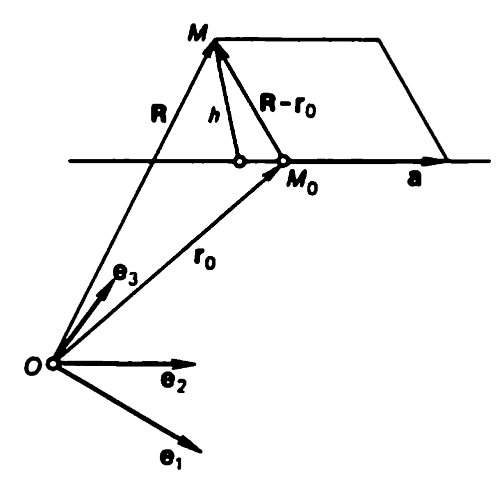


Fig. 24.

Nous omettons d'écrire cette expression en coordonnées si la droite est définie dans l'espace. Considérons une droite dans le plan. Soit $Ax + By + C = 0$ son équation dans le repère cartésien rectangulaire. Prenons pour vecteur directeur le vecteur $\mathbf{a}(-B, A)$. Selon la formule (15), § 3, ch. I,

$$h = \frac{|(X - x_0)A - (Y - y_0)(-B)|}{\sqrt{A^2 + B^2}}, \tag{11}$$

où $M_0(x_0, y_0)$ est un point de la droite, X, Y les composantes de $\mathbf{R}$. Ayant en vue que $C = -Ax_0 - By_0$, on obtient

$$h = \frac{|AX + BY + C|}{\sqrt{A^2 + B^2}}. \tag{12}$$

En partant de (11) on remarque aisément que

$$h = \frac{|(\mathbf{R} - \mathbf{r}_0, \mathbf{n})|}{|\mathbf{n}|},$$

où $\mathbf{n}(A, B)$ est le vecteur normal de la droite. En comparant cette expression à la formule (8), on voit que l'analogie entre plans et droites dans le plan se conserve également dans ce problème.

L'équation de la forme

$$\frac{Ax + By + C}{\sqrt{A^2 + B^2}} = 0$$

s'appelle *équation normée de la droite* dans le plan. S'y rapporte une proposition analogue à la proposition 2.

8. Distance entre deux droites non parallèles dans l'espace. Soient p et q deux droites non parallèles. On sait que dans ce cas il existe deux plans

parallèles P et Q tels que la droite p appartient au plan P et la droite q au plan Q. (Si les équations des droites sont $\mathbf{r} - \mathbf{r}_1 = t\mathbf{a}_1$ et $\mathbf{r} - \mathbf{r}_2 = t\mathbf{a}_2$, le plan P a pour point initial $\mathbf{r}_1$ et pour vecteurs directeurs $\mathbf{a}_1$ et $\mathbf{a}_2$. De façon analogue se construit le plan Q.) La distance h entre les plans P et Q est la *distance entre les droites* p et q. Si p et q se coupent, P et Q se confondent et $h = 0$.

Pour calculer la distance h, le plus simple est de diviser le volume du parallélépipède porté par les vecteurs $\mathbf{r}_2 - \mathbf{r}_1$, $\mathbf{a}_1$ et $\mathbf{a}_2$ par l'aire de sa base (fig. 25). Il vient

$$h = \frac{|(\mathbf{r}_2 - \mathbf{r}_1, \mathbf{a}_1, \mathbf{a}_2)|}{|[\mathbf{a}_1, \mathbf{a}_2]|}.$$

Le dénominateur de cette expression est différent de zéro vu que les droites ne sont pas parallèles.

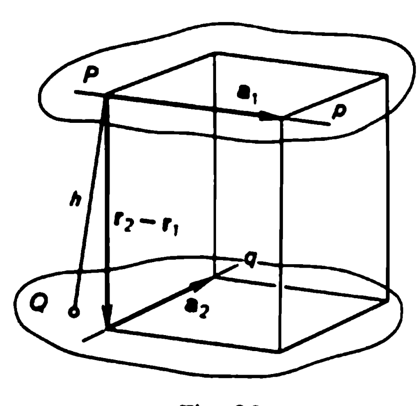


Fig. 25.

Proposition 3. *Deux droites d'équations* $\mathbf{r} = \mathbf{r}_1 + \mathbf{a}_1 t$ *et* $\mathbf{r} = \mathbf{r}_2 + \mathbf{a}_2 t$ *se coupent si et seulement si* $h = 0$, *c'est-à-dire si*

$$(\mathbf{r}_2 - \mathbf{r}_1, \mathbf{a}_1, \mathbf{a}_2) = 0, \quad [\mathbf{a}_1, \mathbf{a}_2] \neq 0.$$

9. Calcul des angles. Pour déterminer l'angle de deux droites, il faut rechercher leurs vecteurs directeurs et calculer l'angle qu'ils forment d'après la formule (4) du § 3, ch. I. Ceci étant, il faut avoir en vue qu'en choisissant, sur l'une des droites, le vecteur directeur de sens contraire, on obtient un angle adjacent supplémentaire.

Pour déterminer l'angle entre la droite et le plan, on définit l'angle θ entre le vecteur directeur de la droite et le vecteur perpendiculaire au plan. Si le vecteur directeur de la droite est choisi de sorte que $\cos\theta \geqslant 0$, et $0 \leqslant \theta \leqslant \pi/2$, l'angle entre la droite et le plan est le complémentaire de θ.

L'angle de deux plans est celui des vecteurs perpendiculaires aux plans.

Les calculs se font le plus aisément dans le repère cartésien rectangulaire. Pour vecteur perpendiculaire au plan on peut alors choisir le vecteur

de composantes égales aux coefficients associés aux variables de l'équation de ce plan.

Supposons que deux droites du plan sont définies dans le repère cartésien rectangulaire par les équations

$$y = k_1 x + b_1, \quad y = k_2 x + b_2. \tag{13}$$

Désignons par φ l'angle mesuré de la première droite à la seconde dans le sens de la plus petite rotation amenant le premier vecteur de base sur le second ; tg φ est la tangente de la différence des angles que les droites forment avec l'axe des abscisses. Or les tangentes de ces derniers sont égales aux coefficients angulaires des droites. Il vient alors

$$\operatorname{tg} \varphi = \frac{k_2 - k_1}{1 + k_1 k_2}. \tag{14}$$

Il faut souligner que cette formule perd son sens quand le dénominateur s'annule. Dans ce cas les droites sont perpendiculaires. En effet, selon la proposition 1 du § 2, les vecteurs de composantes $(1, k_1)$ et $(1, k_2)$ sont les vecteurs directeurs des droites et leur produit scalaire est $1 + k_1 k_2$.

PROPOSITION 4. *Pour que les droites* (13) *soient perpendiculaires, il faut et il suffit que soit remplie l'égalité*

$$1 + k_1 k_2 = 0.$$

10. Signification géométrique de l'ordre d'une courbe algébrique. Soit donnée dans le plan une courbe algébrique L dont l'équation dans le repère cartésien est

$$A_1 x^{k_1} y^{l_1} + A_2 x^{k_2} y^{l_2} + \ldots + A_s x^{k_s} y^{l_s} = 0. \tag{15}$$

Considérons dans le même repère une droite définie par les équations paramétriques

$$x = x_0 + a_1 t, \quad y = y_0 + a_2 t. \tag{16}$$

Cherchons les points d'intersections de la droite avec la courbe L. Ils seront connus si l'on trouve les valeurs correspondantes du paramètre t. Ce seront les valeurs pour lesquelles x et y exprimés par les formules (16) vérifieront l'équation (15). Portons (16) dans (15), il vient

$$A_1 (x_0 + a_1 t)^{k_1} (y_0 + a_2 t)^{l_1} + \ldots$$
$$\ldots + A_s (x_0 + a_1 t)^{k_s} (y_0 + a_2 t)^{l_s} = 0. \tag{17}$$

En chassant les parenthèses dans chaque terme, on obtient des polynômes en t de degrés $k_1 + l_1, \ldots, k_s + l_s$. Leur somme est un polynôme en t de degré au plus égal au degré maximal des termes. Or le plus grand des nombres $k_1 + l_1, \ldots, k_s + l_s$ est l'ordre de la courbe L. Donc, le degré de l'équation (17) ne dépasse pas l'ordre de la courbe.

Il se peut évidemment que tous les coefficients de l'équation soient égaux à zéro et cette dernière constitue une identité. Dans ce cas tous les points de la droite appartiennent à la courbe. Si l'on exclut ce cas, le nombre de racines de l'équation, et, partant, de points d'intersection, ne dépasse pas l'ordre de la courbe. On a ainsi démontré la proposition suivante.

PROPOSITION 5. *Le nombre de points d'intersection de la courbe algébrique L avec la droite qui n'y appartient pas toute entière ne peut dépasser l'ordre de la courbe L.*

Il existe des courbes dont le nombre maximal de points d'intersection avec une droite n'est jamais égal à l'ordre de la courbe. C'est le cas par exemple de la courbe d'ordre 4 dont l'équation dans le repère cartésien rectangulaire est $y = x^4$. Elle coupe toute droite en deux points au plus (fig. 26).

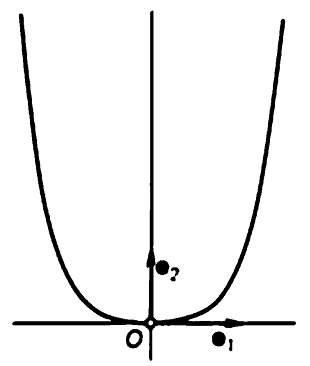


Fig. 26. Courbe d'équation $y = x^4$

EXEMPLE. La spirale d'Archimède est la courbe d'équation $r = a\varphi$ dans le système de coordonnées polaires ; chaque droite passant par le pôle la coupe en un nombre infini de points. Elle n'est donc pas une courbe algébrique.

CHAPITRE III

CONIQUES ET QUADRIQUES

Dans les deux premiers paragraphes de ce chapitre, on traitera des courbes planes du deuxième ordre appelées aussi coniques. Une telle courbe peut être définie par une équation du second degré, de forme générale *)

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0, \tag{1}$$

où les coefficients A, B et C ne sont pas simultanément nuls.

§ 1. Etude de l'équation du second degré

Proposons-nous d'étudier une conique. A cet effet, considérons l'équation (1) dans laquelle les coefficients A, B et C ne s'annulent pas simultanément. Cherchons un ensemble de points vérifiant l'équation (1), sans savoir à priori si au moins un de ces points existe. Pour cela, on procédera aux changements du repère cartésien afin de rendre l'équation (1) aussi simple que possible et on admettra dès le début que le repère cartésien est rectangulaire, vu que le passage à un repère rectangulaire ne fait pas varier la forme générale de l'équation (1).

Si on fait tourner le repère cartésien rectangulaire de l'angle φ, les anciennes coordonnées x, y deviennent liées aux nouvelles coordonnées x', y' par les formules de passage (voir (7), § 4, ch. I)

$$x = x'\cos\varphi - y'\sin\varphi, \quad y = x'\sin\varphi + y'\cos\varphi.$$

Dans les nouvelles coordonnées, l'équation (1) prend la forme

$$A(x'\cos\varphi - y'\sin\varphi)^2 + 2B(x'\cos\varphi - y'\sin\varphi) \times$$
$$\times (x'\sin\varphi + y'\cos\varphi) + C(x'\sin\varphi + y'\cos\varphi)^2 + \ldots = 0.$$

Les points de suspension désignent ici les termes du premier degré en x' et y', ainsi que le terme constant, que nous n'avons pas besoin d'expliciter.

*) Les coefficients affectant le produit des variables x et y et leurs puissances premières sont notés $2B$, $2D$ et $2E$, vu que plus loin on aura affaire aux moitiés de ces coefficients.

On s'intéressera dans l'équation transformée au terme qui contient le produit $x'y'$. On peut calculer aisément que son coefficient est

$$B' = -A\sin\varphi\cos\varphi + B(\cos^2\varphi - \sin^2\varphi) + C\sin\varphi\cos\varphi,$$

car les termes non écrits ne renferment pas le produit $x'y'$. Si $B = 0$, on ne fera pas tourner le repère. Si $B \neq 0$, choisissons un angle φ de manière que B' s'annule. Cette exigence implique l'équation $2B\cos 2\varphi = (A - C)\sin 2\varphi$ en φ. Si $A = C$, $\cos 2\varphi = 0$ et l'on peut poser $\varphi = \pi/4$. Mais si $A \neq C$, on choisit $\varphi = \frac{1}{2}$ Arctg $[2B/(A - C)]$. Il est essentiel qu'il existe toujours au moins un angle φ pour lequel B' s'annule. Ainsi, après avoir tourné le repère de l'angle φ, on obtient

$$A'x'^2 + C'y'^2 + 2D'x' + 2E'y' + F' = 0. \qquad (2)$$

Les expressions des coefficients de l'équation (2) en fonction des coefficients de l'équation (1) et de l'angle φ s'obtiennent facilement mais on n'en a pas besoin. L'important est que par rotation du repère l'équation arbitraire du second degré puisse être ramenée à la forme (2). B' est maintenant nul, les autres coefficients étant toujours considérés arbitraires.

Enonçons la proposition auxiliaire suivante.

PROPOSITION 1. *Si le carré d'une des coordonnées figure dans l'équation* (2) *avec coefficient non nul, on peut, par déplacement de l'origine des coordonnées le long de l'axe correspondant, rendre nul le terme contenant la puissance première de cette coordonnée.*

En effet, soit par exemple $A' \neq 0$. Ecrivons (2) sous la forme

$$A'\left(x'^2 + \frac{2D'}{A'}x' + \left(\frac{D'}{A'}\right)^2\right) + C'y'^2 + 2E'y' + F' - \frac{D'^2}{A'} = 0.$$

Si l'on effectue le déplacement de l'origine des coordonnées défini par les formules $x'' = x' + D'/A'$, $y'' = y'$, l'équation (2) prend la forme

$$A'x''^2 + C'y''^2 + 2E'y'' + F'' = 0.$$

La proposition est démontrée.

A) *Supposons que dans l'équation* (2) *on a* $A'C' \neq 0$, c'est-à-dire que A' et C' sont tous les deux différents de zéro. Selon la proposition 1, l'équation peut être mise sous la forme

$$A'x''^2 + C'y''^2 + F'' = 0. \qquad (2')$$

On peut faire les hypothèses suivantes sur les signes des coefficients de cette équation.

1) *Les coefficients A' et C' ont même signe*. Quant au signe de F'', il y a trois éventualités :

a) *Le signe de F'' est opposé à celui de A' et C'.* Faisons passer F'' dans le second membre de l'égalité et divisons-la par F''. L'équation prend la forme

$$\frac{x''^2}{a^2} + \frac{y''^2}{b^2} = 1, \tag{3}$$

où $a^2 = -F''/A'$, $b^2 = -F''/C'$.

On peut admettre que dans cette équation on a $a \geqslant b$. En effet, si $a < < b$, on procède à un nouveau changement de coordonnées

$$x^* = y'', \quad y^* = x''. \tag{4}$$

Définition. La courbe qui dans un repère cartésien rectangulaire est définie par l'équation (3) pour $a \geqslant b$ s'appelle *ellipse*, et l'équation (3) *équation canonique de l'ellipse.*

Pour $a = b$, l'équation (3) est l'*équation du cercle* de rayon a. Donc, le cercle est un cas particulier de l'ellipse.

b) *Le signe de F'' est le même que celui de A' et C'.* Comme précédemment, on peut alors réduire l'équation à la forme

$$\frac{x''^2}{a^2} + \frac{y''^2}{b^2} = -1. \tag{5}$$

Cette équation n'est vérifiée par les coordonnées d'aucun point. L'équation réduite à la forme canonique (5) est appelée *équation de l'ellipse imaginaire.*

c) $F'' = 0$. L'équation prend la forme

$$a^2x''^2 + c^2y''^2 = 0. \tag{6}$$

Elle est vérifiée par un seul point $x'' = 0, y'' = 0$. L'équation réduite à la forme canonique (6) s'appelle *équation d'un couple de droites sécantes imaginaires.* La raison de cette appellation est puisée dans la ressemblance avec l'équation (8) produite plus bas.

2) *Les coefficients A' et C' sont de signes opposées.* En ce qui concerne le terme constant, deux éventualités se présentent.

a) $F'' \neq 0$. Si nécessaire, en effectuant la substitution (4), on peut admettre que le signe de F'' est opposé à celui de A'. L'équation se réduit alors à la forme

$$\frac{x''^2}{a^2} - \frac{y''^2}{b^2} = 1, \tag{7}$$

où $a^2 = -F''/A'$, $b^2 = F''/C'$.

Définition. La courbe qui dans un repère cartésien rectangulaire peut être définie par l'équation (7) s'appelle *hyperbole*, et l'équation (7) *équation canonique de l'hyperbole.*

b) $F'' = 0$. L'équation prend la forme

$$a^2x''^2 - c^2y''^2 = 0. \tag{8}$$

Son premier membre se décompose en facteurs

$$(ax'' - cy'')(ax'' + cy'')$$

et, par suite, s'annule si et seulement si l'un au moins des facteurs est nul. Aussi la courbe d'équation (8) est-elle constituée de deux droites. Ces droites se coupent à l'origine des coordonnées et on a ainsi un couple de droites sécantes.

B) *Admettons maintenant que* $A'C' = 0$ et que, par suite, l'un des coefficients A' ou C' est nul. En cas de nécessité, en effectuant la substitution (4), on peut considérer que $A' = 0$. Signalons que $C' \neq 0$, car autrement le degré de l'équation (2) diminuerait. En se servant de la proposition 1 on peut réduire l'équation de la courbe à la forme

$$C'y''^2 + 2D'x' + F'' = 0.$$

a) *Posons que* $D' \neq 0$. Groupons les termes de la façon suivante :

$$C'y''^2 + 2D'\left(x' + \frac{F''}{2D'}\right) = 0.$$

Maintenant il devient évident qu'en déplaçant l'origine des coordonnées le long de l'axe des abscisses, en accord avec les formules de passage $x^* = x' + F''/2D'$, $y^* = y''$, on réduit l'équation à la forme

$$C'y^{*2} + 2Dx^* = 0$$

ou

$$y^{*2} = 2px^*, \tag{9}$$

avec $p = -D'/C'$. On peut admettre que $p > 0$, sinon on effectue le changement des coordonnées pour modifier le sens de l'axe des abscisses.

DÉFINITION. La courbe qui dans un repère cartésien rectangulaire peut être définie par l'équation (9), avec $p > 0$, s'appelle *parabole*, et l'équation (9) *équation canonique de la parabole*.

b) *Admettons maintenant que* $D' = 0$, c'est-à-dire que l'équation est de la forme

$$C'y''^2 + F'' = 0.$$

Si *les signes de* C' *et* F'' *sont opposés*, en divisant par C' on peut écrire $y''^2 - a^2 \equiv (y'' - a)(y'' + a) = 0$. Chacune des relations $y'' - a = 0$ et $y'' + a = 0$ définit une droite, de sorte que la courbe se réduit à un *couple de droites parallèles*.

Si *les signes de* C' *et* F'' *coïncident*, en divisant par C' on réduit

l'équation à la forme canonique

$$y''^2 + a^2 = 0. \qquad (10)$$

Cette équation n'est vérifiée par aucun point. L'équation réduite à la forme canonique (10) s'appelle *équation d'un couple de droites parallèles imaginaires*.

Il peut arriver que $F'' = 0$. Alors l'équation de la courbe équivaut à $y''^2 = 0$, si bien que la conique représente une droite. Si donc l'équation (1) se réduit à la forme canonique $y''^2 = 0$, son premier membre est le carré d'un polynôme linéaire, de sorte que l'équation (1) est équivalente à une équation linéaire. Cette équation est appelée *équation d'un couple de droites confondues*.

Rassemblons les résultats obtenus.

THÉORÈME 1. *Soit donnée, dans un repère cartésien, l'équation du second degré*

$$Ax^2 + 2Bxy + Cy^2 + 2Dx + 2Ey + F = 0.$$

Il existe un repère cartésien rectangulaire dans lequel cette équation prend l'une des neuf formes canoniques suivantes :

1) $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = 1$, 2) $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} = -1$,

3) $a^2x^2 + c^2y^2 = 0$, 4) $\dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} = 1$,

5) $a^2x^2 - c^2y^2 = 0$, 6) $y^2 = 2px$,

7) $y^2 - a^2 = 0$, 8) $y^2 + a^2 = 0$,

9) $y^2 = 0$.

Il existe donc sept classes de coniques : 1) ellipses, 3) points (couples de droites sécantes imaginaires), 4) hyperboles, 5) couples de droites sécantes, 6) paraboles, 7) couples de droites parallèles, 9) droites (couples de droites confondues).

A l'équation 2) de l'ellipse imaginaire et à l'équation 8) du couple de droites parallèles imaginaires il ne correspond aucun point.

§ 2. Ellipse, hyperbole et parabole

Dans le paragraphe précédent, on a classifié les coniques. Il n'y a que trois classes de ces courbes dont les propriétés géométriques ne sont pas évidentes. On abordera leur étude dans le présent paragraphe.

1. Ellipse. Rappelons que l'ellipse est par définition une courbe qui

dans un repère cartésien rectangulaire est définie par l'équation canonique

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \tag{1}$$

avec $a \geqslant b$. Le repère dont il s'agit dans la définition est appelé *repère canonique*.

Il découle immédiatement de (1) que pour tous les points de l'ellipse on a $|x| \leqslant a$ et $|y| \leqslant b$, c'est-à-dire que l'ellipse s'inscrit dans le rectangle de côtés $2a$ et $2b$. Les points d'intersection de l'ellipse avec les axes du repère canonique, dont les coordonnées sont $(a, 0)$, $(-a, 0)$, $(0, b)$ et $(0, -b)$, sont appelés *sommets* de l'ellipse. Les distances a et b entre l'origine des coordonnées et les sommets s'appellent respectivement *demi-grand axe* et *demi-petit axe* de l'ellipse.

L'équation canonique (1) ne contenant que les carrés des coordonnées possède la propriété suivante : si elle est vérifiée par les coordonnées (x, y) d'un point M, elle l'est encore par les coordonnées $(-x, y)$, $(x, -y)$ et $(-x, -y)$ des points M_1, M_2 et M_3 (fig. 27). D'où la

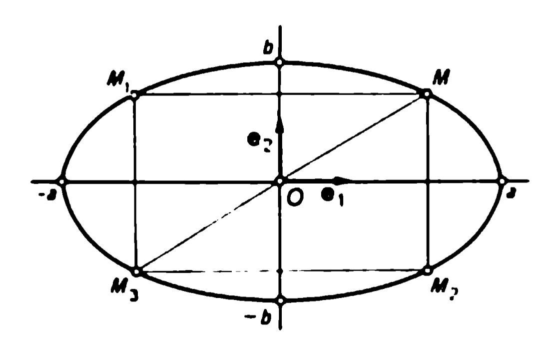


Fig. 27.

PROPOSITION 1. *Les axes du repère canonique sont les axes de symétrie de l'ellipse et l'origine du repère canonique est son centre de symétrie.*

Le centre de symétrie de l'ellipse est appelé tout simplement son *centre*.

Pour décrire la forme géométrique de l'ellipse (1), on la compare le plus souvent au cercle dont le rayon est a et dont le centre se confond avec le centre de l'ellipse. Ecrivons l'équation de ce cercle sous la forme

$$\frac{x^2}{a^2} + \frac{y^2}{a^2} = 1.$$

Pour tout x tel que $|x| < a$ il existe deux points du cercle d'ordonnées $\pm a\sqrt{1 - x^2/a^2}$ et deux points de l'ellipse d'ordonnées $\pm b\sqrt{1 - x^2/a^2}$. Supposons qu'à tout point du cercle correspond un point de l'ellipse dont l'ordonnée est de même signe. Alors le rapport des ordonnées des points

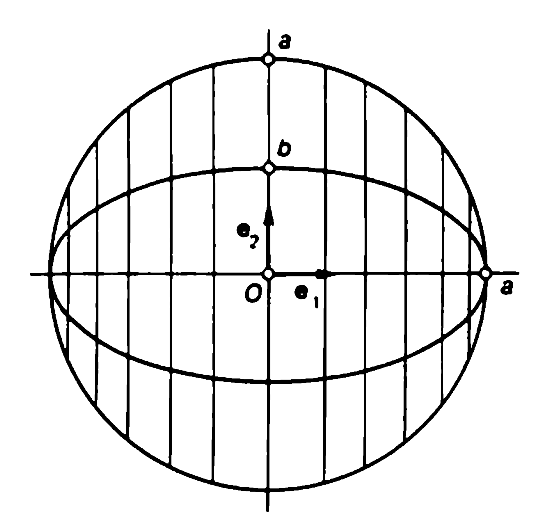


Fig. 28. Ellipse obtenue par contraction du cercle pour $b/a = 1/2$

correspondants est égal à b/a et on peut donc obtenir l'ellipse à partir du cercle par une contraction dans laquelle l'ordonnée de chaque point du cercle diminue dans un même rapport b/a (fig. 28).

Pour effectuer une transformation géométrique du cercle en l'ellipse, prenons un deuxième plan P_2 qui coupe le plan initial P_1 suivant l'axe des abscisses sous l'angle $\alpha = \text{Arccos}\,(b/a)$ (fig. 29). Choisissons dans le plan P_2 un repère cartésien rectangulaire $\{O, \mathbf{e}_1, \mathbf{e}_2'\}$ dont l'origine et le vecteur de base $\mathbf{e}_1$ se confondent avec l'origine et le vecteur $\mathbf{e}_1$ du repère choisi dans le plan P_1, et le vecteur $\mathbf{e}_2'$ forme l'angle α avec le vecteur $\mathbf{e}_2$. Considérons le cercle $X^2 + Y^2 = a^2$ dans le plan P_2. Si du point $M(X, Y)$ de P_2 on abaisse la perpendiculaire sur le plan P_1, les coordonnées (x, y) de son pied N s'obtiennent par les formules $x = X$, $y = Y \cos\alpha = bY/a$. Aussi l'ellipse est-elle l'ensemble des pieds des perpendiculaires abaissées des points du cercle ou, comme on dit, est la *projection orthogonale du cercle.*

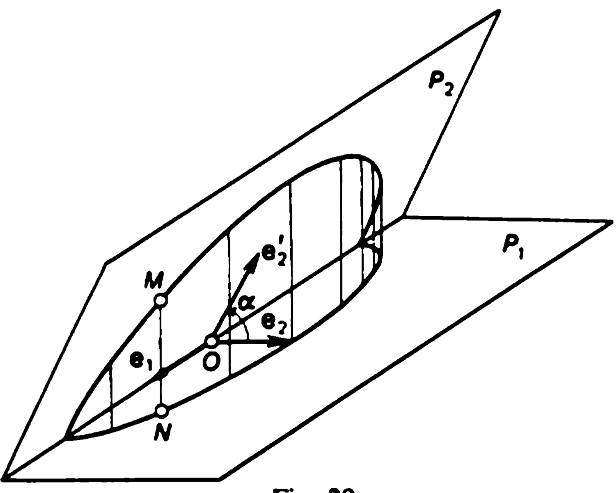


Fig. 29.

L'ellipse admet deux points remarquables appelés ses *foyers*. Soit par définition

$$c^2 = a^2 - b^2 \tag{2}$$

et $c \geqslant 0$. On appelle *foyers* les points F_1 et F_2 dont les coordonnées par rapport au repère canonique sont respectivement $(c, 0)$ et $(-c, 0)$.

Pour le cercle, $c = 0$ et les deux foyers se confondent au centre. On admettra plus bas que l'ellipse n'est pas un cercle.

Le rapport

$$\varepsilon = \frac{c}{a} \tag{3}$$

est appelé *excentricité* de l'ellipse. Signalons qu'on a toujours $\varepsilon < 1$.

PROPOSITION 2. *La distance d'un point arbitraire $M(x, y)$ de l'ellipse à chacun de ses foyers est une fonction linéaire de son abscisse x :*

$$\left.\begin{aligned} r_1 &= |F_1M| = a - \varepsilon x, \\ r_2 &= |F_2M| = a + \varepsilon x. \end{aligned}\right\} \tag{4}$$

DÉMONSTRATION. Il va de soi que

$$r_1 = \sqrt{(x-c)^2 + y^2}$$

(fig. 30). En y portant l'expression de y^2 tirée de l'équation de l'ellipse, on obtient

$$r_1 = \sqrt{x^2 - 2cx + c^2 + b^2 - \frac{b^2x^2}{a^2}}.$$

Transformons l'expression sous le radical en tenant compte de l'égalité (2). Il vient

$$r_1 = \sqrt{a^2 - 2cx + \frac{c^2x^2}{a^2}}.$$

On voit que sous le radical se trouve le carré d'un binôme linéaire, c'est-à-dire que $r_1 = |a - \varepsilon x|$. Comme $\varepsilon < 1$ et $x \leqslant a$, on en déduit que $a - \varepsilon x >$ > 0. On a démontré la première des égalités (4). La seconde se démontre de façon analogue.

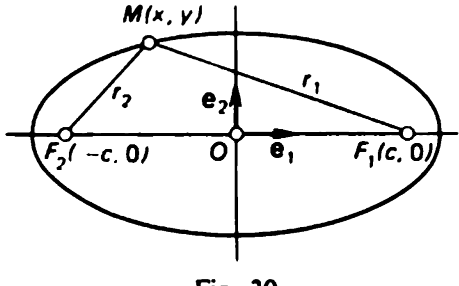


Fig. 30.

PROPOSITION 3. *Pour qu'un point se trouve sur l'ellipse il faut et il suffit que la somme de ses distances aux foyers soit égale à* $2a$, *c'est-à-dire au grand axe de l'ellipse.*

La nécessité de la condition est évidente : si l'on additionne les égalités (4) membre à membre, on obtient

$$r_1 + r_2 = 2a. \tag{5}$$

Démontrons que la condition est suffisante. Supposons que la condition (5) est satisfaite pour le point $M(x, y)$, c'est-à-dire que

$$\sqrt{(x-c)^2 + y^2} = 2a - \sqrt{(x+c)^2 + y^2}.$$

Elevons les deux membres de l'égalité au carré et réduisons les termes semblables :

$$xc + a^2 = a\sqrt{(x+c)^2 + y^2}. \tag{6}$$

Elevons également au carré cette égalité et réduisons les termes semblables en nous servant de la relation (2). On aboutit ainsi à l'égalité équivalente à (1).

L'ellipse admet également deux droites remarquables appelées ses *directrices*. Leurs équations dans le repère canonique sont (fig. 31) :

$$x = \frac{a}{\varepsilon} \quad \text{et} \quad x = -\frac{a}{\varepsilon}. \tag{7}$$

La directrice et le foyer qui se trouvent d'un côté du centre sont considérés comme correspondants l'une à l'autre.

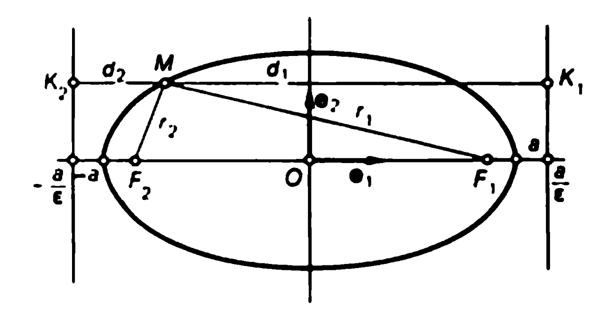


Fig. 31.

PROPOSITION 4. *Pour qu'un point se trouve sur l'ellipse il faut et il suffit que le rapport de ses distances au foyer et à la directrice correspondante soit égal à l'excentricité de l'ellipse* ε.

Démontrons la proposition pour le foyer $F_2(-c, 0)$. Notons d_2 la distance d'un point quelconque de l'ellipse $M(x, y)$ à la directrice d'équation

$x = -a/\varepsilon$. Alors selon la formule (12) du § 3, ch. II, on a

$$d_2 = x + \frac{a}{\varepsilon} = \frac{1}{\varepsilon}(\varepsilon x + a),$$

ce qui diffère de l'expression (4) pour r_2 par un seul facteur $1/\varepsilon$.

Inversement, supposons que pour un point quelconque du plan on a $r_2/d_2 = \varepsilon$, c'est-à-dire que

$$\frac{\sqrt{(x+c)^2 + y^2}}{x + a/\varepsilon} = \varepsilon.$$

Compte tenu de ce que $\varepsilon = c/a$, on réduit aisément cette égalité à la forme (6), de laquelle il résulte, comme on le sait, l'équation de l'ellipse (1).

Pour le second foyer, la proposition découle de la symétrie de l'ellipse par rapport à l'axe des ordonnées du repère canonique.

Cherchons maintenant l'équation de la tangente à l'ellipse définie par l'équation canonique. On sait que le coefficient angulaire de la droite tangente au graphique d'une fonction en un point (x_0, y_0) est égal à la dérivée de cette fonction au point x_0. Posons $y_0 \neq 0$ et considérons la fonction $f(x)$ dont le graphique est entièrement situé sur l'ellipse et passe par le point (x_0, y_0). (Pour $y_0 > 0$, c'est la fonction $f_1(x) = b\sqrt{1 - x^2/a^2}$, pour $y_0 < 0$, la fonction $f_2(x) = -b\sqrt{1 - x^2/a^2}$. Sans préciser le signe de y_0, notons $f(x)$ la fonction qui convient.) La fonction $f(x)$ vérifie l'identité

$$\frac{x^2}{a^2} + \frac{(f(x))^2}{b^2} = 1.$$

En la dérivant par rapport à x, on obtient

$$\frac{2x}{a^2} + \frac{2ff'}{b^2} = 0.$$

Portons $x = x_0$, $f(x_0) = y_0$ dans l'égalité obtenue et résolvons-la par rapport à $f'(x_0)$. On obtient en vertu de $y_0 \neq 0$

$$f'(x_0) = -\frac{b^2}{a^2}\frac{x_0}{y_0}.$$

On est maintenant en mesure d'écrire l'équation de la tangente au point (x_0, y_0) :

$$y - y_0 = -\frac{b^2}{a^2}\frac{x_0}{y_0}(x - x_0).$$

Pour simplifier cette équation, transformons-la sous la forme $a^2yy_0 + b^2xx_0 = b^2x_0^2 + a^2y_0^2$. Vu que le point (x_0, y_0) vérifie l'équation de

l'ellipse, on a $b^2x_0^2 + a^2y_0^2 = a^2b^2$, d'où $a^2yy_0 + b^2xx_0 = a^2b^2$. Ainsi, l'équation de la tangente à l'ellipse au point (x_0, y_0) est de la forme

$$\frac{xx_0}{a^2} + \frac{yy_0}{b^2} = 1. \tag{8}$$

En déduisant l'équation (8), on a exclu les sommets de l'ellipse $(a, 0)$ et $(-a, 0)$ en posant $y_0 \neq 0$. Pour ces points l'équation (8) se transforme respectivement en équations $x = a$ et $x = -a$ qui définissent les tangentes aux sommets. On peut le vérifier en remarquant que x, considéré comme une fonction de y, atteint son extrémum aux sommets. Laissons au lecteur le soin d'élucider la question en détail et montrer que l'équation (8) définit la tangente en tout point (x_0, y_0) de l'ellipse.

PROPOSITION 5. *La tangente à l'ellipse au point $M_0(x_0, y_0)$ est la bissectrice de l'angle adjacent supplémentaire de l'angle formé par les segments joignant ce point aux foyers.*

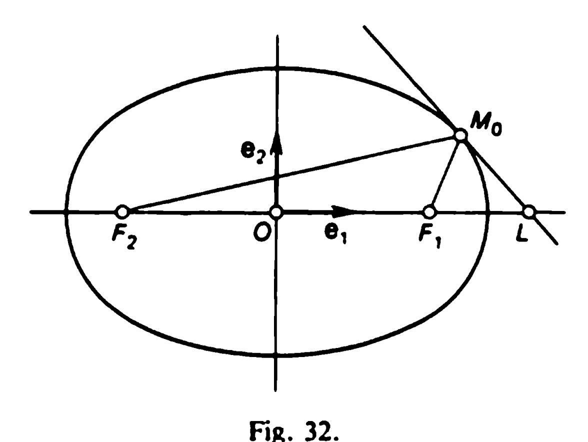


Fig. 32.

DÉMONSTRATION. Soit L le point d'intersection de la tangente avec l'axe des abscisses (fig. 32). De l'équation (8) on tire aussitôt que son abscisse est a^2/x_0. Remarquons que $a^2/|x_0| > a$ et, partant, L se trouve à l'extérieur du segment F_1F_2. Les distances de L aux foyers sont égales à $|F_1L| = |a^2/x_0 - c|$ et $|F_2L| = |a^2/x_0 + c|$ et par suite, leur rapport

$$\frac{|F_1L|}{|F_2L|} = \frac{|a^2/x_0 - c|}{|a^2/x_0 + c|} = \frac{|a - \varepsilon x_0|}{|a + \varepsilon x_0|}$$

est égal à celui des longueurs des segments M_0F_1 et M_0F_2. Il s'ensuit que M_0L est la bissectrice de l'angle extérieur du triangle $M_0F_1F_2$, ce qu'il fallait démontrer. Il faut remarquer que cette démonstration perd son sens pour les sommets de l'ellipse. Mais pour ces derniers l'assertion est évidente.

2. Hyperbole. On a appelé *hyperbole* une courbe qui dans un repère cartésien rectangulaire est définie par l'équation canonique

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \tag{9}$$

Le repère dont il s'agit dans la définition est dit *canonique*.

Il découle immédiatement de l'équation (9) que pour tous les points de l'hyperbole on a $|x| \geqslant a$, c'est-à-dire que tous les points de l'hyperbole se trouvent en dehors de la bande verticale large de $2a$ (fig. 33). L'axe des abscisses du repère canonique coupe l'hyperbole en deux points $(a, 0)$ et $(-a, 0)$ appelés *sommets* de l'hyperbole. L'axe des ordonnées ne coupe pas l'hyperbole. Les nombres a et b sont respectivement appelés *demi-axe transverse* et *demi-axe non transverse* de l'hyperbole.

Exactement comme pour l'ellipse on démontre la proposition suivante.

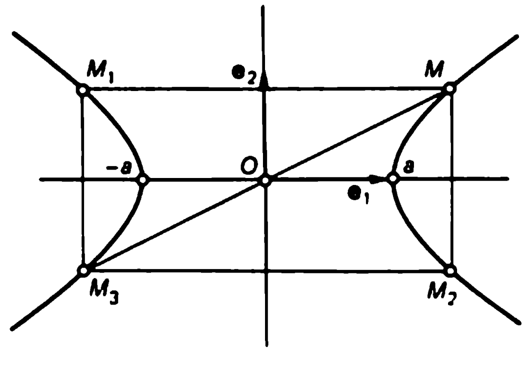


Fig. 33.

PROPOSITION 6. *Les axes du repère canonique sont les axes de symétrie de l'hyperbole, et l'origine du repère canonique, son centre de symétrie.*

Le centre de symétrie de l'hyperbole est appelé son *centre* (fig. 33).

Pour étudier la forme de l'hyperbole, cherchons son intersection avec une droite arbitraire passant par l'origine des coordonnées. Soit $y = kx$ l'équation de la droite, car on sait déjà que la droite $x = 0$ ne coupe pas l'hyperbole. Les abscisses des points d'intersection s'obtiennent à partir de l'équation

$$\frac{x^2}{a^2} - \frac{k^2x^2}{b^2} = 1,$$

ou si $b^2 - a^2k^2 > 0$,

$$x = \pm\frac{ab}{\sqrt{b^2 - a^2k^2}}.$$

On est ainsi en mesure d'indiquer les coordonnées de deux points d'intersection :

$$(ab/v, abk/v) \quad \text{et} \quad (-ab/v, -abk/v),$$

où $v = (b^2 - a^2k^2)^{1/2}$. En vertu de la symétrie, il suffit de suivre le mouvement du premier point avec la variation de k (fig. 34).

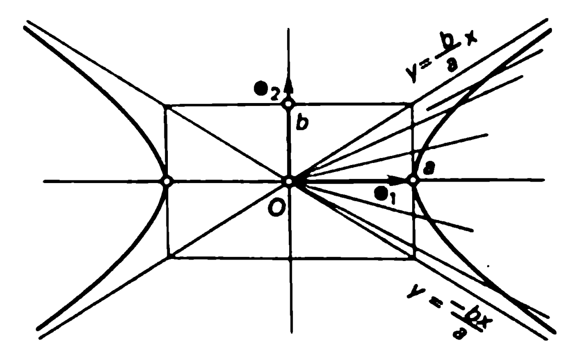


Fig. 34.

Le numérateur de la fraction ab/v est constant, tandis que le dénominateur présente sa plus grande valeur pour $k = 0$. Par suite, le point de coordonnées $(a, 0)$ possède la plus petite abscisse. Avec l'accroissement de k le dénominateur diminue et l'abscisse x augmente en tendant vers l'infini quand k s'approche du nombre b/a. L'hyperbole ne rencontre pas la droite $y = bx/a$ de coefficient angulaire b/a et *a fortiori*, elle n'est coupée par aucune droite de coefficient angulaire plus grand que b/a. Mais toute droite de coefficient angulaire positif inférieur à b/a coupe l'hyperbole.

Si on commence à tourner la droite à partir de la position horizontale dans le sens des aiguilles d'une montre, son coefficient k diminue, tandis que k^2 augmente et la droite coupe l'hyperbole en des points qui s'éloignent de plus en plus, jusqu'à ce que le coefficient de la droite ne devienne égal à $-b/a$. A la droite $y = -bx/a$ se rapporte tout ce qui a été dit à propos de $y = bx/a$: elle ne coupe pas l'hyperbole et sépare les droites qui la coupent de celles qui ne la coupent pas.

Définition. Les droites définies dans un repère canonique par les équations $y = bx/a$ et $y = -bx/a$ sont appelées *asymptotes* de l'hyperbole.

Des raisonnements faits plus haut il découle que la représentation géométrique de l'hyperbole est celle de la figure 34. L'hyperbole est composée de deux parties appelées ses *branches*.

Ecrivons les équations des asymptotes sous la forme $bx - ay = 0$ et $bx + ay = 0$. Les distances du point M de coordonnées (x, y) aux asymptotes sont respectivement égales à (comp. (12), § 3, ch. II)

$$h_1 = \frac{|bx - ay|}{\sqrt{a^2 + b^2}}, \quad h_2 = \frac{|bx + ay|}{\sqrt{a^2 + b^2}}.$$

Si le point M est sur l'hyperbole, on a $b^2x^2 - a^2y^2 = a^2b^2$ et

$$h_1h_2 = \frac{|b^2x^2 - a^2y^2|}{a^2 + b^2} = \frac{a^2b^2}{a^2 + b^2}.$$

PROPOSITION 7. *Le produit des distances d'un point de l'hyperbole aux asymptotes est constant et vaut* $\dfrac{a^2b^2}{a^2 + b^2}$.

D'où découle une importante propriété des asymptotes.

PROPOSITION 8. *Si un point se déplace sur l'hyperbole de manière que son abscisse augmente indéfiniment en valeur absolue, la distance de ce point à l'une des asymptotes tend vers zéro.*

En effet, au moins une des distances h_1 ou h_2 doit dans ces conditions croître indéfiniment, et si la proposition était fausse, le produit ne pourrait être constant.

L'hyperbole admet deux points remarquables appelés ses *foyers*. Introduisons un nombre c en posant par définition

$$c^2 = a^2 + b^2 \qquad (10)$$

et $c > 0$. On appelle *foyers* de l'hyperbole dans le repère canonique les points F_1 et F_2 de coordonnées respectivement $(c, 0)$ et $(-c, 0)$ (fig. 35).

Par analogie à l'ellipse, on appelle *excentricité* de l'hyperbole le rapport $\varepsilon = c/a$. Pour l'hyperbole on a toujours $\varepsilon > 1$.

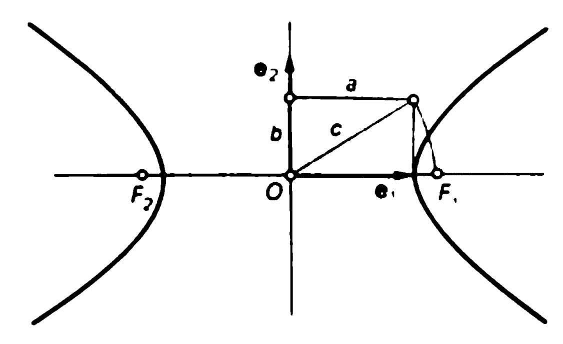


Fig. 35.

PROPOSITION 9. *Les distances d'un point arbitraire $M(x, y)$ de l'hyperbole à chacun des foyers dépendent de l'abscisse x de la façon suivante :*

$$\begin{aligned} r_1 &= |F_1M| = |a - \varepsilon x|, \\ r_2 &= |F_2M| = |a + \varepsilon x|. \end{aligned} \qquad (11)$$

La démonstration de cette proposition reproduit mot à mot celle de la proposition 2 et l'on ne la répétera pas.

Les égalités (11) peuvent être écrites d'une manière plus détaillée

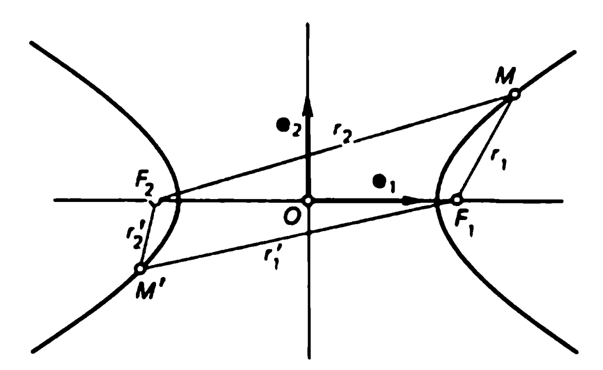


Fig. 36. $r_2 - r_1 = 2a$, $r'_1 - r'_2 = 2a$

(fig. 36) : pour $x \geqslant a$ (branche droite de l'hyperbole)

$$r_1 = \varepsilon x - a, \quad r_2 = \varepsilon x + a \;;$$

pour $x \leqslant -a$ (branche gauche de l'hyperbole)

$$r_1 = a - \varepsilon x, \quad r_2 = -\varepsilon x - a.$$

On voit que pour la branche droite de l'hyperbole on a $r_2 - r_1 = 2a$ et pour la gauche, $r_1 - r_2 = 2a$. Dans les deux cas

$$|r_1 - r_2| = 2a. \tag{12}$$

Ceci démontre la nécessité de la condition énoncée dans la proposition suivante.

PROPOSITION 10. *Pour qu'un point appartienne à l'hyperbole il faut et il suffit que la valeur absolue de la différence de ses distances aux foyers soit égale à l'axe transverse de l'hyperbole.*

Pour démontrer que la condition est suffisante, on doit la mettre sous la forme

$$\pm\sqrt{(x-c)^2+y^2} = 2a \pm \sqrt{(x+c)^2+y^2}.$$

Dans la suite, la démonstration s'écarte de celle de la proposition 3 par le seul fait qu'on se sert de l'égalité (10) au lieu d'utiliser l'égalité (2).

A l'hyperbole se rattachent deux droites appelées ses *directrices*. Leurs équations dans le repère canonique sont :

$$x = \frac{a}{\varepsilon} \quad \text{et} \quad x = -\frac{a}{\varepsilon}. \tag{13}$$

Les directrices de l'hyperbole sont plus proches du centre que ses sommets et, par suite, ne coupent pas l'hyperbole. La directrice et le foyer se trouvant du même côté du centre seront considérés comme correspondants l'une à l'autre.

Proposition 11. *Pour qu'un point se trouve sur l'hyperbole il faut et il suffit que le rapport de ses distances au foyer et à la directrice correspondante soit égal à l'excentricité ε de l'hyperbole.*

Démontrons cette proposition pour le foyer F_2. Désignons par d_2 la distance d'un point arbitraire de l'hyperbole $M(x, y)$ à la directrice d'équation $x = -a/\varepsilon$ (fig. 37). Alors, selon la formule (12) du § 3, ch. II, il vient

$$d_2 = \left| x + \frac{a}{\varepsilon} \right| = \frac{1}{\varepsilon} \, |\varepsilon x + a|,$$

ce qui diffère de r_2 (comp. (11)) par le seul facteur $1/\varepsilon$.

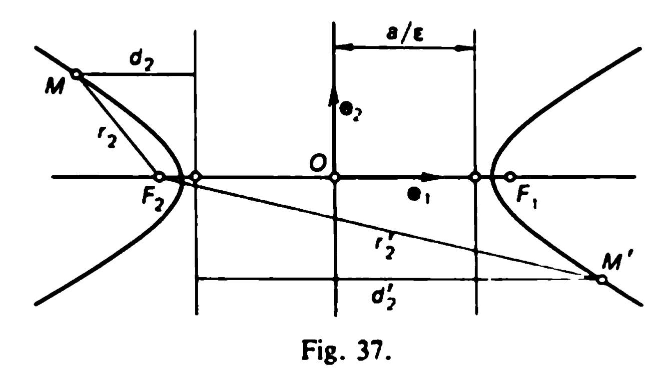


Fig. 37.

La suffisance de la condition se démontre de la même façon que pour la proposition 4 se rapportant à l'ellipse.

L'équation de la tangente en un point (x_0, y_0) de l'hyperbole se déduit de la même façon que l'équation correspondante (8) pour l'ellipse. Elle est de la forme

$$\frac{xx_0}{a^2} - \frac{yy_0}{b^2} = 1. \tag{14}$$

Proposition 12. *La tangente à l'hyperbole au point $M_0(x_0, y_0)$ est la bissectrice de l'angle formé par les segments joignant ce point aux foyers.*

La démonstration ne s'écarte presque pas de celle de la proposition 5. On recommande au lecteur de démontrer cette proposition ainsi que toutes les autres assertions sur l'hyperbole énoncées dans ce point mais non démontrées.

3. Parabole. Rappelons que la parabole est par définition une courbe qui, dans un repère cartésien rectangulaire, est définie par l'équation canonique

$$y^2 = 2px, \tag{15}$$

à la condition que $p > 0$. Le repère dont il s'agit dans la définition est dit *canonique*.

Il ressort de l'équation (15) que pour tous les points de la parabole on a $x \geqslant 0$. La parabole passe par l'origine du repère canonique. Ce point est appelé *sommet* de la parabole.

L'équation de la parabole étant en même temps vérifiée par les points $M(x, y)$ et $M_1(x, -y)$, l'axe des abscisses du repère canonique est un axe de symétrie de la parabole.

La forme de la parabole est bien connue depuis l'école secondaire en tant que graphique de la fonction $y = ax^2$. La différence entre les équations est due au fait que par rapport au repère précédent on a changé de place les axes du repère canonique et désigné le coefficient a^{-1} par $2p$.

On appelle *foyer* F de la parabole rapportée au repère canonique le point de coordonnées $(p/2, 0)$ (fig. 38).

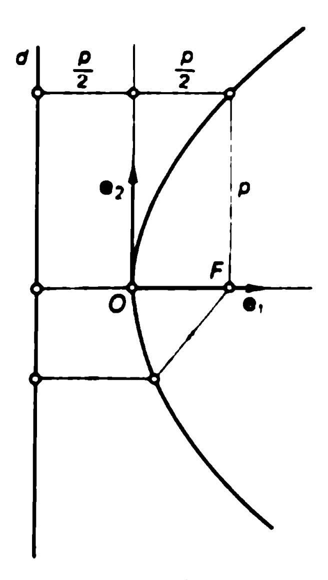


Fig. 38.

La *directrice* de la parabole est la droite d'équation $x = -p/2$ dans le repère canonique.

PROPOSITION 13. *La distance d'un point arbitraire de la parabole au foyer vaut*

$$r = x + \frac{p}{2}\,. \tag{16}$$

Pour le démontrer, portons y^2 de l'équation (15) dans l'expression $r = \sqrt{(x - p/2)^2 + y^2}$ de la distance du point $M(x, y)$ au foyer :

$$r = \sqrt{\left(x - \frac{p}{2}\right)^2 + px}.$$

En transformant le radicande, on obtient

$$r = \left| x + \frac{p}{2} \right|,$$

ce qui entraîne (16) en vertu de $x \geqslant 0$.

Remarquons que d'après la formule (12) du § 3, ch. II, la distance du point de la parabole à la directrice est aussi égale à

$$d = x + \frac{p}{2},$$

d'où la nécessité de la condition énoncée dans la proposition suivante.

PROPOSITION 14. *Pour qu'un point M appartienne à la parabole il faut et il suffit qu'il soit équidistant du foyer et de la directrice de cette parabole.*

Démontrons que la condition est suffisante. Soit un point $M(x, y)$ équidistant du foyer et de la directrice de la parabole (15), c'est-à-dire que

$$\sqrt{\left(x - \frac{p}{2}\right)^2 + y^2} = x + \frac{p}{2}.$$

En élevant cette égalité au carré et en réduisant les termes semblables, on en déduit l'équation de la parabole. La démonstration est donc achevée.

On attribue à la parabole l'excentricité $\varepsilon = 1$. En vertu de cette convention, la formule

$$\frac{r}{d} = \varepsilon$$

qui relie les distances du point de la courbe au foyer et à la directrice, sera également vraie pour l'ellipse, pour l'hyperbole et pour la parabole.

Déduisons l'équation de la tangente en un point $M_0(x_0, y_0)$ de la parabole. Soit $y_0 \neq 0$. Considérons la fonction $y = f(x)$ dont le graphique est entièrement situé sur la parabole et passe par le point M_0. (C'est $y = \sqrt{2px}$ ou $y = -\sqrt{2px}$, suivant le signe de y_0.) La fonction $f(x)$ vérifie l'identité $(f(x))^2 = 2px$, dont la dérivation donne $2f(x)f'(x) = 2p$. En substituant $x = x_0$ et $f(x_0) = y_0$, on obtient

$$f'(x_0) = \frac{p}{y_0}$$

puisque $y_0 \neq 0$. Maintenant on peut écrire l'équation de la tangente à la parabole :

$$y - y_0 = \frac{p}{y_0}(x - x_0).$$

Pour la simplifier, chassons les parenthèses : $yy_0 - y_0^2 = px - px_0$, et signalons que $y_0^2 = 2px_0$. Maintenant l'équation de la tangente à la parabole prend sa forme définitive :

$$yy_0 = p(x + x_0). \tag{17}$$

Remarquons que pour le sommet de la parabole qu'on a exclu en posant $y_0 \neq 0$, l'équation (17) se transforme en l'équation $x = 0$, c'est-à-dire en l'équation de la tangente au sommet. Ainsi donc, l'équation (17) est vérifiée pour tout point $M_0(x_0, y_0)$ de la parabole.

PROPOSITION 15. *La tangente à la parabole au point M_0 est la bissectrice d'un angle adjacent supplémentaire de l'angle formé par le segment joignant le point M_0 au foyer et par la demi-droite issue de ce point en direction de l'axe de la parabole* (fig. 39).

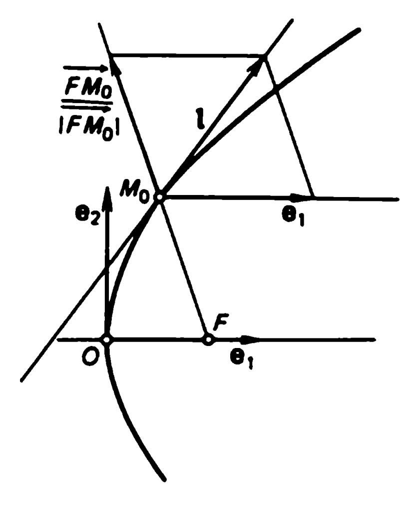


Fig. 39.

DÉMONSTRATION. Le vecteur **l** dirigé le long de la bissectrice de l'angle mentionné dans la proposition est la somme de deux vecteurs unités portés par les côtés de l'angle, c'est-à-dire de $\overrightarrow{FM_0}/|\overrightarrow{FM_0}|$ et de $\mathbf{e}_1$. Si les coordonnées de M_0 sont x_0 et y_0, $\overrightarrow{FM_0} = (x_0 - p/2)\mathbf{e}_1 + y_0\mathbf{e}_2$ et $|\overrightarrow{FM_0}| = x_0 + p/2$. Donc,

$$\mathbf{l} = \left(\frac{x_0 - p/2}{x_0 + p/2} + 1\right)\mathbf{e}_1 + \frac{y_0}{x_0 + p/2}\mathbf{e}_2 = \frac{1}{x_0 + p/2}(2x_0\mathbf{e}_1 + y_0\mathbf{e}_2).$$

On peut maintenant calculer le coefficient angulaire de la bissectrice : $k = y_0/(2x_0)$. On a calculé plus haut le coefficient angulaire de la tangente : $k' = p/y_0$. Montrons que $k = k'$. En effet, $k/k' = y_0^2/(2px_0) = 1$. La proposition est ainsi démontrée.

§ 3. Quadriques

Par analogie au § 2 où on a décrit les coniques les plus intéressantes, on traitera dans le présent paragraphe les quadriques (surfaces du deuxième ordre) les plus importantes, en reléguant au chapitre IX leur classification plus complète. Pour se représenter de façon générale la plupart des quadriques, on peut considérer les surfaces engendrées par rotation des coniques autour de leurs axes de symétrie.

1. Surfaces de révolution. On dit que la surface S est *surface de révolution* d'axe d si elle se compose de cercles situés dans des plans perpendiculaires à la droite d et dont les centres appartiennent à cette droite. Cette définition peut être interprétée de la façon suivante. Considérons une courbe L située dans un plan P qui passe par l'axe de révolution d (fig. 40) et faisons-la tourner autour de cet axe. Chaque point de la courbe décrit alors un cercle, et toute la courbe, une surface de révolution.

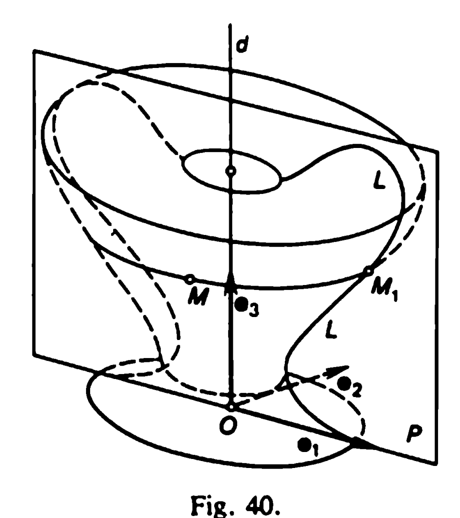


Fig. 40.

Choisissons un repère cartésien rectangulaire $\{O, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ de telle sorte que son origine appartienne à l'axe de révolution d, le vecteur $\mathbf{e}_3$ soit dirigé le long de d et le vecteur $\mathbf{e}_1$ situé dans le plan P. Ainsi donc, le point O et les vecteurs $\mathbf{e}_1$ et $\mathbf{e}_3$ constituent un repère cartésien dans le plan P. Supposons que la courbe L dont la rotation engendre une surface soit définie dans ce repère par l'équation $\varphi(x, z) = 0$.

Considérons un point $M(x, y, z)$. Par ce point passe un cercle de centre sur l'axe d et qui se trouve dans le plan perpendiculaire à l'axe. Le rayon du cercle est égal à la distance séparant M de l'axe, c'est-à-dire à $\sqrt{x^2 + y^2}$. Le point M se trouve sur la surface de révolution si et seulement si sur le cercle mentionné il existe un point M_1 appartenant à la courbe L.

Le point $M_1(x_1, y_1, z_1)$ se trouve dans le plan P, de sorte que $y_1 = 0$. En outre, $z_1 = z$ et $|x_1| = \sqrt{x^2 + y^2}$, car M_1 appartient au cercle passant

par M. Les coordonnées du point M_1 vérifient l'équation de la courbe $L : \varphi(x_1, z_1) = 0$. En portant x_1 et z_1 dans cette équation, on obtient la condition suivante pour les coordonnées du point M

$$\varphi(\pm\sqrt{x^2 + y^2}, z) = 0, \tag{1}$$

qui est nécessaire et suffisante pour que le point M se trouve sur la surface de révolution S (l'égalité (1) doit être vérifiée pour l'un au moins des deux signes précédant la racine). Cette condition peut être écrite sous une forme équivalente

$$\varphi(\sqrt{x^2 + y^2}, z)\varphi(-\sqrt{x^2 + y^2}, z) = 0 \tag{2}$$

et constitue justement l'équation de la surface de révolution de la courbe L autour de l'axe d.

2. Ellipsoïde. Considérons les surfaces engendrées par rotation d'une ellipse autour de ses axes de symétrie. En dirigeant le vecteur $\mathbf{e}_3$ d'abord le long du petit axe et ensuite le long du grand axe, on obtient l'équation de l'ellipse sous les formes suivantes

$$\frac{x^2}{a^2} + \frac{z^2}{c^2} = 1, \quad \frac{z^2}{a^2} + \frac{x^2}{c^2} = 1$$

(c désigne ici le demi-petit axe de l'ellipse). En vertu de la formule (1), les équations des surfaces de révolution respectives seront

$$\frac{x^2 + y^2}{a^2} + \frac{z^2}{c^2} = 1 \tag{3}$$

et

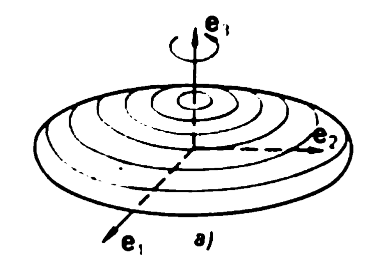
$$\frac{z^2}{a^2} + \frac{x^2 + y^2}{c^2} = 1. \tag{4}$$

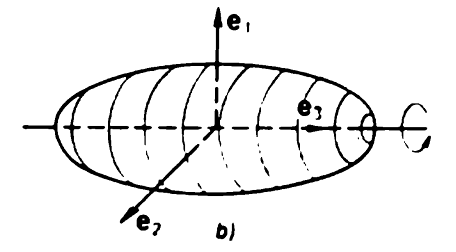
Les quadriques (3) et (4) sont appelées *ellipsoïdes de révolution aplati* et *allongé*. On les a représentés sur la figure 41 *a*, *b*.

Déplaçons chaque point $M(x, y, z)$ de l'ellipsoïde de révolution (3) vers le plan Y (plan de coordonnées passant par $\mathbf{e}_1$ et $\mathbf{e}_3$) de manière que la distance du point à ce plan diminue dans un rapport $\lambda < 1$ constant pour tous les points. A la suite de ce déplacement, le point M vient se confondre avec le point M' dont les coordonnées sont définies par les égalités $x' = x, y' = \lambda y, z' = z$. Ainsi, tous les points de l'ellipsoïde de révolution (3) deviennent des points de la surface d'équation

$$\frac{x'^2}{a^2} + \frac{y'^2}{b^2} + \frac{z'^2}{c^2} = 1, \tag{5}$$

où $b = \lambda a$. La quadrique qui dans un repère cartésien rectangulaire a pour équation (5) est appelée *ellipsoïde* (voir fig. 41, *c*). S'il arrive que $b = c$, on obtient encore un ellipsoïde de révolution mais, cette fois, allongé.





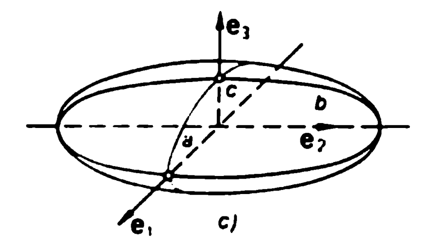


Fig. 41. a), b) Ellipsoïdes de révolution aplati et allongé ; c) ellipsoïde

L'ellipsoïde, tout comme l'ellipsoïde de révolution à partir duquel il a été obtenu, représente une surface bornée fermée. Il découle de l'équation (5) que l'ellipsoïde est symétrique par rapport à l'origine et par rapport aux plans de coordonnées.

En effectuant une contraction, on arrive à obtenir un ellipsoïde à partir de l'ellipsoïde de révolution, tout comme on obtient une ellipse à partir du cercle. Par contraction de la sphère $x^2 + y^2 + z^2 = a^2$, on peut obtenir l'ellipsoïde de révolution (3). Pour obtenir un ellipsoïde allongé à partir de la sphère, il faut effectuer une transformation analogue, cette fois avec $\lambda > 1$, c'est-à-dire une traction.

On reviendra à maintes reprises dans ce paragraphe à la contraction sans toutefois la décrire chaque fois en détail.

3. Cône d'ordre 2. Considérons dans le plan P un couple de droites sécantes, défini dans le repère $\{O, \mathbf{e}_1, \mathbf{e}_3\}$ par l'équation $a^2x^2 - c^2z^2 = 0$. La surface (fig. 42) engendrée par rotation de ces droites autour de l'axe des cotes a pour équation

$$a^2(x^2 + y^2) - c^2z^2 = 0 \tag{6}$$

et porte le nom de *cône circulaire droit*. La contraction vers le plan Y trans-

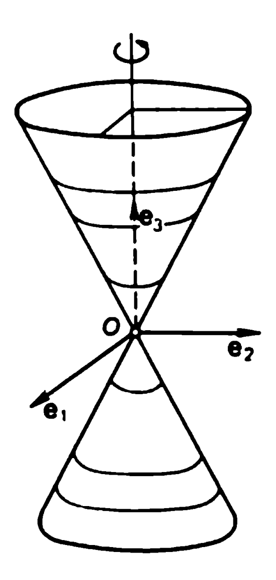


Fig. 42.

forme le cône circulaire droit en une surface dont l'équation est

$$a^2x^2 + b^2y^2 - c^2z^2 = 0. \tag{7}$$

La quadrique qui dans un repère cartésien rectangulaire a pour équation (7) est appelée *cône* ou, d'une façon plus précise, *cône d'ordre* 2. Le cône est composé de droites passant par l'origine des coordonnées. Les sections du cône par les plans $z = \alpha$ sont, pour différents α, des ellipses

$$a^2x^2 + b^2y^2 = c^2\alpha^2.$$

4. Hyperboloïde à une nappe. Un *hyperboloïde de révolution à une nappe* est une surface engendrée par rotation de l'hyperbole

$$\frac{x^2}{a^2} - \frac{z^2}{c^2} = 1$$

autour de son axe non transverse. La formule (1) permet d'obtenir l'équation de l'hyperboloïde de révolution à une nappe

$$\frac{x^2 + y^2}{a^2} - \frac{z^2}{c^2} = 1. \tag{8}$$

En contractant cette surface, on obtient une quadrique d'équation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1. \tag{9}$$

La quadrique qui, dans un repère cartésien rectangulaire, est définie par l'équation (9) s'appelle *hyperboloïde à une nappe* (fig. 43).

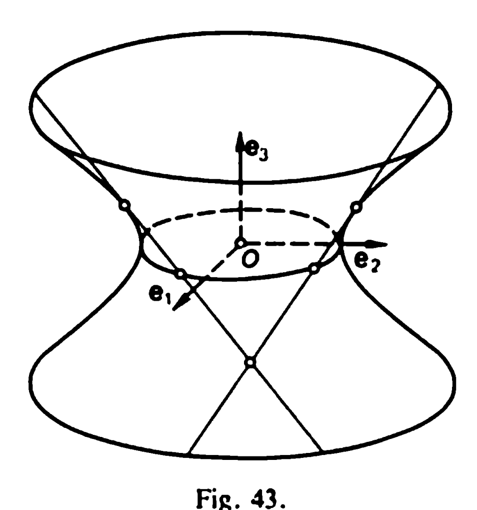


Fig. 43.

Notons une propriété intéressante de cette surface : l'hyperboloïde à une nappe admet des *génératrices rectilignes*. On appelle ainsi les droites dont tous les points se trouvent sur la surface. Par tout point de l'hyperboloïde à une nappe passent deux génératrices rectilignes dont les équations peuvent être obtenues de la façon suivante. Portons le terme y^2/b^2 dans le second membre de l'équation (9) et décomposons les deux membres de l'égalité en facteurs :

$$\left(\frac{x}{a}+\frac{z}{c}\right)\left(\frac{x}{a}-\frac{z}{c}\right)=\left(1+\frac{y}{b}\right)\left(1-\frac{y}{b}\right).$$

Considérons maintenant une droite définie par les équations

$$\left.\begin{aligned}\mu\left(\frac{x}{a}+\frac{z}{c}\right)&=\lambda\left(1+\frac{y}{b}\right),\\ \lambda\left(\frac{x}{a}-\frac{z}{c}\right)&=\mu\left(1-\frac{y}{b}\right),\end{aligned}\right\}\qquad(10)$$

où λ et μ sont des nombres. Les coordonnées de chaque point de cette droite doivent satisfaire à deux équations et, par suite, à leur produit, c'est-à-dire à l'équation (9). Ainsi donc, quels que soient λ et μ, tous les points de la droite d'équations (10) se trouvent sur l'hyperboloïde à une nappe. Le même raisonnement peut aussi être produit pour la famille de droites

$$\left.\begin{aligned}\lambda'\left(\frac{x}{a}+\frac{z}{c}\right)&=\mu'\left(1-\frac{y}{b}\right),\\ \mu'\left(\frac{x}{a}-\frac{z}{c}\right)&=\lambda'\left(1+\frac{y}{b}\right).\end{aligned}\right\}\qquad(11)$$

En portant les coordonnées du point de l'hyperboloïde à une nappe dans l'une des équations (10) et l'une des équations (11), on obtient les valeurs des paramètres λ, μ et λ', μ' correspondant aux génératrices rectilignes qui passent par ce point. Il est naturel que chaque couple de paramètres soit défini à un facteur commun près.

Si on fait tourner l'hyperbole avec ses asymptotes, ces dernières décriront un cône circulaire droit appelé *cône asymptote* de l'hyperboloïde de révolution. La contraction de l'hyperboloïde de révolution en un hyperboloïde à une nappe entraîne celle du cône circulaire droit en un cône appelé *cône asymptote* de l'hyperboloïde à une nappe.

5. Hyperboloïde à deux nappes. Un *hyperboloïde de révolution à deux nappes* est une surface engendrée par rotation de l'hyperbole

$$\frac{z^2}{c^2} - \frac{x^2}{a^2} = 1$$

autour de son axe transverse. En vertu de la formule (1), l'équation de l'hyperboloïde de révolution à deux nappes est de la forme

$$\frac{z^2}{c^2} - \frac{x^2 + y^2}{a^2} = 1. \tag{12}$$

En contractant cette surface, on obtient une quadrique d'équation

$$\frac{z^2}{c^2} - \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1. \tag{13}$$

La quadrique définie dans un repère cartésien rectangulaire par une équation de la forme (13) s'appelle *hyperboloïde à deux nappes* (fig. 44). A

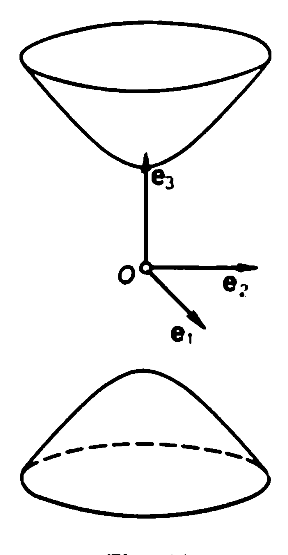


Fig. 44.

deux branches de l'hyperbole correspondent ici deux parties de la surface (nappes) tandis que, dans le cas de l'hyperboloïde de révolution à une nappe, chaque branche de l'hyperbole engendre la surface entière.

Le cône asymptote se définit pour l'hyperboloïde à deux nappes de la même façon que pour l'hyperboloïde à une nappe.

6. Paraboloïde elliptique. En faisant tourner la parabole $x^2 = 2pz$ autour de son axe de symétrie, on obtient une surface d'équation

$$x^2 + y^2 = 2pz, \tag{14}$$

appelée *paraboloïde de révolution*. La contraction vers le plan $y = 0$ transforme le paraboloïde de révolution en une surface d'équation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 2z. \tag{15}$$

La quadrique définie par une telle équation dans un repère cartésien rectangulaire est appelée *paraboloïde elliptique* (fig. 45). Sa forme géométrique

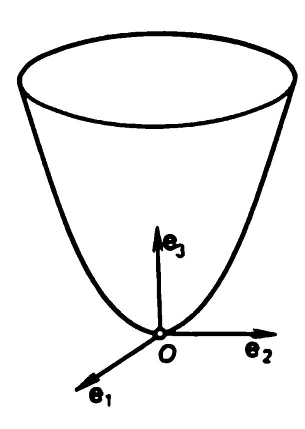


Fig. 45.

est bien évidente. Signalons que les sections de cette surface par les plans $z = \alpha$, avec $\alpha > 0$, sont les ellipses

$$\frac{x^2}{2\alpha a^2} + \frac{y^2}{2\alpha b^2} = 1, \quad z = \alpha,$$

tandis que les sections par des plans parallèles à d'autres plans de coordonnées, par exemple par les plans $y = \alpha$, représentent les paraboles

$$\frac{x^2}{a^2} + \frac{\alpha^2}{b^2} = 2z, \quad y = \alpha.$$

7. Paraboloïde hyperbolique. Par analogie avec l'équation (15), on peut écrire l'équation

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 2z. \tag{16}$$

La quadrique qui dans un repère cartésien rectangulaire prend la forme de l'équation (16) sera appelée *paraboloïde hyperbolique*. Etudions la forme géométrique de cette surface. Pour cela, considérons la section du paraboloïde hyperbolique par le plan $x = \alpha$ pour un α quelconque. Choisissons dans ce plan un repère cartésien rectangulaire $\{O', \mathbf{e}_2, \mathbf{e}_3\}$ d'origine au point $O'(\alpha, 0, 0)$. Rapportée à ce repère, la ligne d'intersection a pour équation

$$-\frac{y^2}{b^2} = 2\left(z - \frac{\alpha^2}{2a^2}\right). \tag{17}$$

Cette courbe est une parabole. On s'en assure facilement en reportant l'origine des coordonnées au point O'' de coordonnées $(0, \alpha^2/2a^2)$ (ses coordonnées dans l'espace par rapport au repère initial $\{O, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ sont $(\alpha, 0, \alpha^2/2a^2)$). Le point O'' est évidemment le sommet de la parabole, l'axe de la parabole est parallèle au vecteur $\mathbf{e}_3$, tandis que le signe moins dans le premier membre de l'égalité (17) signifie que les branches de la parabole sont dirigées dans le sens opposé à celui du vecteur $\mathbf{e}_3$. Signalons qu'après le transport de l'origine des coordonnées au point O'', l'équation de la parabole ne contient plus de α et, partant, toutes les sections du paraboloïde hyperbolique par les plans $x = \alpha$ sont des paraboles égales.

Faisons varier α et suivons le déplacement du sommet O'' de la parabole en fonction de α. Les coordonnées $(\alpha, 0, \alpha^2/2a^2)$ du point O'' par rapport au repère $\{O, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ permettent de conclure que ce point parcourt une courbe définie dans le repère $\{O, \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ par les équations

$$z = \frac{x^2}{2a^2}, \quad y = 0.$$

Cette courbe est une parabole du plan $y = 0$. Son sommet est à l'origine des coordonnées, son axe de symétrie se confond avec l'axe des cotes et ses branches sont de même sens que le vecteur $\mathbf{e}_3$.

On peut maintenant construire un paraboloïde hyperbolique de la façon suivante. Etant donné deux paraboles, déplaçons l'une d'elles de manière que : son sommet glisse suivant l'autre parabole, leurs axes demeurent parallèles, les paraboles se trouvent dans des plans perpendiculaires et leurs branches soient dirigées dans les sens opposés. Animée de ce mouvement, la parabole engendre le paraboloïde hyperbolique (fig. 46).

La section du paraboloïde hyperbolique par le plan $z = \alpha$ est une hyperbole dont l'équation, par rapport au repère $\{O^*, \mathbf{e}_1, \mathbf{e}_2\}$ de ce plan

Fig. 46. *AOB* — parabole immobile, *KLM*, *NOP* et *GRS* — différentes positions de la parabole mobile

avec origine au point $O^*(0, 0, \alpha)$, est de la forme

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 2\alpha.$$

Pour des grands α positifs, les demi-axes de l'hyperbole $\sqrt{2\alpha}\,a$ et $\sqrt{2\alpha}\,b$ sont aussi grands, ils diminuent lorsque α décroît. Ceci étant, l'axe transverse de l'hyperbole est parallèle au vecteur $\mathbf{e}_1$ (fig. 47). Pour $\alpha = 0$, l'hyperbole dégénère en un couple de droites sécantes. Si $\alpha < 0$, l'axe transverse de l'hyperbole devient parallèle au vecteur $\mathbf{e}_2$. Les demi-axes augmentent avec $|\alpha|$. Le rapport des demi-axes de tous les hyperboles est le même si les α sont de même signe. Il s'ensuit que les projections, sur un même plan, de toutes les sections du paraboloïde hyperbolique forment une famille d'hyperboles ayant pour asymptotes le couple de droites sécantes d'équation $\frac{x^2}{a^2} - \frac{y^2}{b^2} = 0$ (fig. 48).

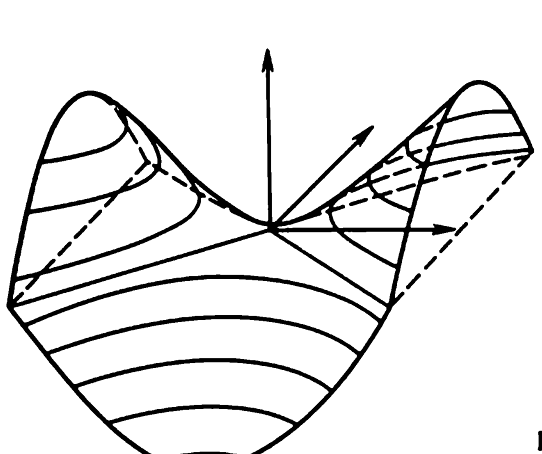


Fig. 47.

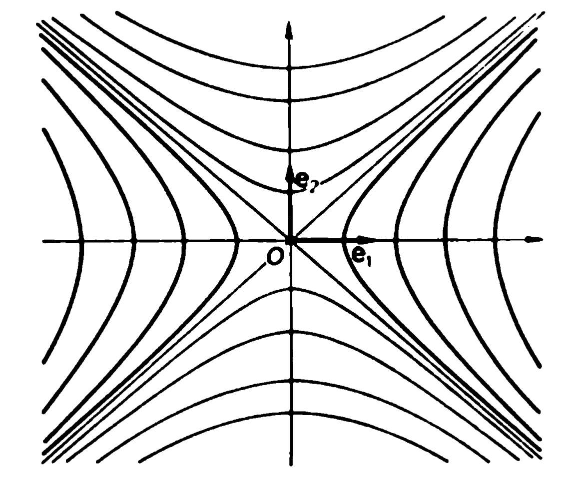


Fig. 48. Projections des sections d'un paraboloïde hyperbolique. Les traits fins correspondent à $\alpha > 0$, les traits forts, à $\alpha > 0$

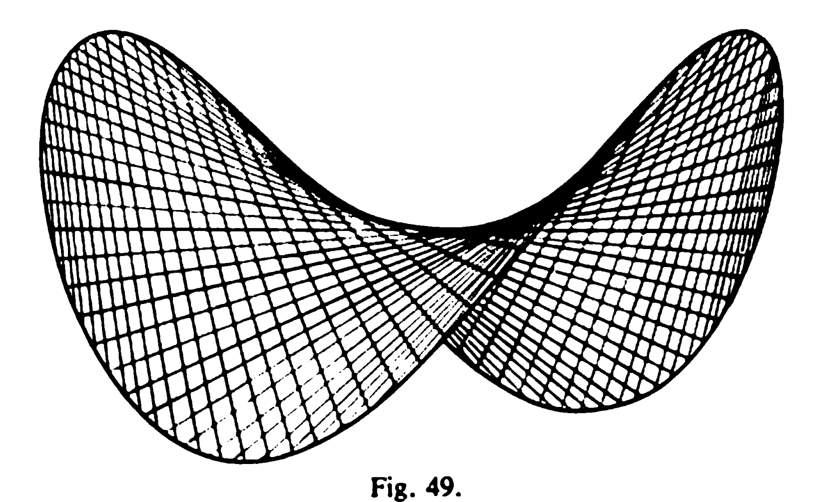


Fig. 49.

Tout comme l'hyperboloïde à une nappe, le paraboloïde hyperbolique possède deux familles de génératrices rectilignes (fig. 49) dont les équations sont de la forme :

$$1)\ \lambda\left(\frac{x}{a}-\frac{y}{b}\right)=\mu\,,\qquad \mu\left(\frac{x}{a}+\frac{y}{b}\right)=2\lambda z,$$

$$2)\ \lambda'\left(\frac{x}{a}+\frac{y}{b}\right)=\mu',\qquad \mu'\left(\frac{x}{a}-\frac{y}{b}\right)=2\lambda' z.$$

Ces équations se déduisent de la même façon que les équations des génératrices rectilignes de l'hyperboloïde à une nappe.

CHAPITRE IV

TRANSFORMATIONS DU PLAN

§ 1. Applications et transformations

1. Définition. On appelle *application* $\mathfrak{f}$ du plan P dans le plan R la relation dans laquelle à tout point du plan P est associé un point déterminé du plan R. On utilisera la notation $\mathfrak{f} : P \to R$. S'il est nécessaire d'indiquer qu'au point A du plan P correspond un point B du plan R, on écrira $B = \mathfrak{f}(A)$. On dit que le point B du plan R est l'*image* de A par l'application $\mathfrak{f}$ et que le point A du plan P est un ***antécédent*** ou une ***image réciproque*** de B.

Les applications pour lesquelles les plans P et R se confondent sont appelées *transformations*. On traitera dans ce paragraphe les faits principaux sur les applications que le lecteur connaît déjà à partir des transformations.

Soulignons que pour les applications, comme pour les transformations, on ne suppose pas que tout point du plan R est une image d'un point de P. Il se peut que l'ensemble de toutes les images ne coïncide pas avec R.

2. Exemples. 1) Etant donné deux plans P et R, faisons correspondre à chaque point du plan P le pied de la perpendiculaire abaissée de ce point sur le plan R. On a ainsi défini une application qu'on appelle *projecteur orthogonal.* Dans cette application, tout point du plan R possède, en général, un antécédent et un seul. Il existe un cas où le projecteur orthogonal change brusquement de caractère. A savoir, si les plans P et R sont perpendiculaires, tout point du plan R n'a pas d'antécédent ; seuls les points de la droite d'intersection des plans en possèdent. En revanche, chacun de ces points a une infinité d'antécédents situés sur la perpendiculaire à R élevée en ce point.

2) La translation, la rotation, la symétrie axiale, l'homothétie, que le lecteur connaît déjà, peuvent servir d'exemples de transformations.

3) Soit la droite p du plan P et soit le nombre $\lambda > 0$. D'un point arbitraire M non situé sur p abaissons la perpendiculaire sur cette droite et désignons son pied par N. Définissons le point $\mathfrak{f}(M)$ par la relation $\overrightarrow{N\mathfrak{f}(M)} = \lambda\overrightarrow{NM}$. Si M est sur p, posons $\mathfrak{f}(M) = M$ (fig. 50). La transformation ainsi définie s'appelle *contraction vers la droite p* dans le rapport λ. (Si $\lambda > 1$, cette transformation peut être appelée *traction*.)

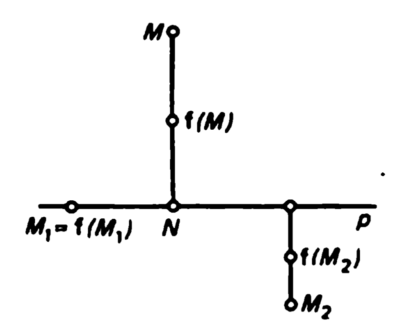


Fig. 50.

On s'est déjà servi de la contraction vers la droite au § 1 du ch. III lors de l'étude de l'ellipse. Une transformation analogue de l'espace, la contraction vers le plan, a été utilisée lors de la définition de la quadrique au § 4 du ch. III.

4) Choisissons dans chacun des plans P et R un repère cartésien rectangulaire et associons à un point de coordonnées x et y dans le plan P un point de coordonnées $x^* = x^2 - y^2, y^* = 2xy$ dans le plan R. On s'assure aisément en résolvant ces équations en x et y que tout point du plan R admet deux antécédents dans le plan P, l'origine des coordonnées faisant exception car son antécédent est unique.

5) Etant donné un point O, associons à tout point M différent de O un point $\mathbf{f}(M)$ défini par l'égalité

$$\overrightarrow{O\mathbf{f}(M)} = \frac{\operatorname{Arctg} |\overrightarrow{OM}|}{|\overrightarrow{OM}|} \overrightarrow{OM}.$$

Posons $\mathbf{f}(O) = O$. Ceci étant, à chaque point du plan est associé un point unique situé à l'intérieur du cercle de rayon $\pi/2$ et de centre au point O. Chaque point intérieur au cercle admet un antécédent et un seul, tandis que les points extérieurs au cercle n'en possèdent pas.

6) On peut associer à tout point du plan le pied de la perpendiculaire abaissée de ce point sur une droite fixée p, et à tout point de la droite ce point même. Dans ce cas, à tous les points de la droite perpendiculaire à p est associé un même point.

7) On peut associer à chaque point du plan P un même point dans le plan R.

3. Produit d'applications. Application réciproque. La notion de composée des transformations est connue dès l'école secondaire. On l'appellera *produit* en la définissant aussi pour les applications.

Définition. Soient les applications $\mathbf{f} : P \to R$ et $\mathbf{g} : R \to S$. L'application $\mathbf{h}$ associant à tout point A du plan P un point $\mathbf{g}(\mathbf{f}(A))$ du plan S est appelée *produit* de l'application $\mathbf{f}$ par l'application $\mathbf{g}$ et est notée $\mathbf{g} \circ \mathbf{f}$. L'application qui agit la première figure à droite.

Soulignons que pour qu'on puisse définir le produit d'applications, il faut que le plan d'arrivée de la première application se confonde avec le plan de départ de la seconde application.

Il va de soi que le produit d'applications, comme celui de transformations, dépend de l'ordre des facteurs, c'est-à-dire que $\mathbf{g} \circ \mathbf{f}$ n'est pas en général égal à $\mathbf{f} \circ \mathbf{g}$. Il faut signaler que pour les applications, les deux produits ne sont définis simultanément que dans le cas où $\mathbf{f} : P \to R$ et $\mathbf{g} : R \to P$.

Laissons au lecteur le soin de se convaincre que la composition des applications est associative, c'est-à-dire que si le produit $(\mathbf{f} \circ \mathbf{g}) \circ \mathbf{h}$ est défini, il en est de même de $\mathbf{f} \circ (\mathbf{g} \circ \mathbf{h})$, et les deux produits sont égaux.

Si l'on note $\mathbf{e}_P$ et $\mathbf{e}_R$ les transformations identiques des plans P et R, toute application $\mathbf{f} : P \to R$ vérifie $\mathbf{f} \circ \mathbf{e}_P = \mathbf{f}$ et $\mathbf{e}_R \circ \mathbf{f} = \mathbf{f}$. Si $\mathbf{f}$ est une transformation du plan, ces égalités se réduisent à $\mathbf{e} \circ \mathbf{f} = \mathbf{f} \circ \mathbf{e} = \mathbf{f}$.

Par définition de l'application $\mathbf{f} : P \to R$, tout point du plan P a une image et une seule. Les exemples 4) et 6) montrent qu'un point du plan R peut avoir plusieurs antécédents, tandis que dans les exemples 5), 6) et 7) tout point du plan R n'est pas nécessairement image d'un point de P.

DÉFINITION. L'application $\mathbf{f} : P \to R$ est appelée *application bijective* ou *bijection* du plan P sur le plan R si tout point du plan R est l'image d'un point du plan P et d'un seul.

Les applications mentionnées dans 2), 3) sont bijectives et dans les exemples 4) à 7) elles ne le sont pas.

Soit donnée l'application $\mathbf{f} : P \to R$. A tout point A du plan P elle associe son image $\mathbf{f}(A)$ dans le plan R. Associons maintenant d'une façon réciproque un point A à tout point $\mathbf{f}(A)$. Il va de soi qu'une telle relation satisfait à notre définition de l'application si et seulement si tout point du plan R est l'image d'un point de P et d'un seul. Cela signifie que l'application $\mathbf{f}$ doit être bijective.

DÉFINITION. On appellera *application réciproque* de l'application bijective $\mathbf{f} : P \to R$ l'application $\mathbf{f}^{-1} : R \to P$ telle que $\mathbf{f}^{-1}(\mathbf{f}(A)) = A$ pour tout point A du plan P.

On voit aisément que la définition de l'application réciproque est équivalente à la relation $\mathbf{f}^{-1} \circ \mathbf{f} = \mathbf{e}_P$, où $\mathbf{e}_P$ est la transformation identique du plan P.

Deux points confondus possèdent une même image, de sorte que $\mathbf{f}(\mathbf{f}^{-1}(\mathbf{f}(A))) = \mathbf{f}(A)$ ou $\mathbf{f}(\mathbf{f}^{-1}(B)) = B$ pour tout point B du plan R. Cela signifie, en particulier, que la réciproque de l'application $\mathbf{f}^{-1}$ est l'application $\mathbf{f}$. La condition $\mathbf{f}(\mathbf{f}^{-1}(B)) = B$ est équivalente à l'égalité $\mathbf{f} \circ \mathbf{f}^{-1} = \mathbf{e}_R$, où $\mathbf{e}_R$ est la transformation identique du plan R.

Si $\mathbf{f}$ est une transformation bijective du plan, on a

$$\mathbf{f}^{-1} \circ \mathbf{f} = \mathbf{f} \circ \mathbf{f}^{-1} = \mathbf{e}.$$

4. Expression analytique d'une application. Soit donnée une application $\mathbf{f} : P \to R$, où P et R sont deux plans. Par définition, cela signifie qu'est donnée une relation dans laquelle à tout point M du plan P est associée son image $\mathbf{f}(M)$ dans le plan R. Si l'on choisit dans le plan P un repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$, et dans le plan R un repère $\{Q, \mathbf{p}_1, \mathbf{p}_2\}$, le point M sera défini par le couple de nombres x, y, et le point M^* par les nombres x^*, y^*. Il s'ensuit que l'application $\mathbf{f}$ fait correspondre à chaque couple de nombres x, y des nombres x^*, y^*. Ainsi donc, étant donné deux repères, définir une application revient à définir deux fonctions dont chacune dépend de deux variables indépendantes :

$$x^* = \varphi(x, y), \qquad y^* = \psi(x, y). \tag{1}$$

Ces égalités représentent une expression analytique, c'est-à-dire avec emploi des coordonnées, qui a été utilisée dans l'exemple 4) de ce paragraphe.

Soulignons que dans le cas général, les repères des plans P et R ne sont pas liés entre eux : le point Q n'est pas obligé d'être l'image du point O, et les vecteurs $\mathbf{p}_1$ et $\mathbf{p}_2$ celles des vecteurs $\mathbf{e}_1$ et $\mathbf{e}_2$.

S'il s'agit d'une transformation, il est tout naturel de choisir un seul repère, vu que l'image et son antécédent se trouvent dans le même plan.

Si les fonctions φ et ψ sont définies pour tout couple de nombres, les formules (1) définissent, dans les repères donnés des plans P et R, une application $\mathbf{f} : P \to R$.

Il s'ensuit, en particulier, que l'étude des applications est un problème aussi vaste que l'étude des fonctions ou des courbes arbitraires. On s'occupera donc ici des applications qui forment une classe restreinte et simple mais fort importante. Parmi toutes les applications elles se détachent de la même façon que les droites parmi toutes les courbes planes.

§ 2. Applications linéaires

1. Définition des applications linéaires. Considérons deux plans P et R.

DÉFINITION. On dit que l'application $\mathbf{f} : P \to R$ est *linéaire* s'il existe des repères cartésiens dans les plans P et R par rapport auxquels $\mathbf{f}$ peut être définie par les formules

$$\left.\begin{aligned} x^* &= a_1x + b_1y + c_1, \\ y^* &= a_2x + b_2y + c_2. \end{aligned}\right\} \tag{1}$$

L'application linéaire bijective est appelée *application affine*.

Soulignons que dans la définition de l'application linéaire il n'est pas du tout exigé que les coefficients a_1 et b_1 ou a_2 et b_2 soient simultanément non nuls. Les seconds membres des formules (1) sont des polynômes de degré

au plus égal à 1. Pour les applications affines ces polynômes sont linéaires et vérifient de plus la

PROPOSITION 1. *Pour qu'une application linéaire définie par les formules* (1) *soit bijective il faut et il suffit que*

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \neq 0. \tag{2}$$

Ainsi, une application affine est définie par les formules (1) *à condition que* (2) *soit vérifiée.*

DÉMONSTRATION. Cette assertion découle de la proposition 10 du § 2, ch. II. En effet, on doit rechercher la condition à laquelle tout point $M^*(x^*, y^*)$ possède un antécédent unique $M(x, y)$. On obtient les coordonnées du point M en résolvant le système (1). Or ce système possède une solution unique, quels que soient les termes constants $x^* - c_1$ et $y^* - c_2$, si et seulement si est satisfaite la condition (2).

Voici quelques exemples d'applications et transformations à la fois affines et linéaires.

1) Le projecteur orthogonal (exemple 1 du § 1) est une application linéaire. Pour le démontrer, choisissons dans les plans P et R deux repères cartésiens rectangulaires de manière que la droite d'intersection de ces plans soit leur axe des abscisses commun. Dans ce cas, étant donné le projecteur orthogonal $\mathfrak{f} : P \to R$, les points M et $\mathfrak{f}(M)$ possèdent la même abscisse, et le rapport de leurs ordonnées est égal au cosinus de l'angle entre les plans. Ainsi donc,

$$x^* = x, \quad y^* = y\cos\varphi.$$

L'application est affine si et seulement si les plans ne sont pas perpendiculaires, c'est-à-dire si $\cos\varphi \neq 0$.

2) L'homothétie s'écrit le plus aisément dans le repère cartésien dont l'origine est le centre d'homothétie O. Puisque tout vecteur $\overrightarrow{OM}$ se transforme par homothétie en un vecteur $\overrightarrow{O\mathfrak{f}(M)} = \lambda\overrightarrow{OM}$, le point M de coordonnées x, y devient le point $\mathfrak{f}(M)$ de coordonnées λx, λy. Ainsi, l'homothétie est définie par les égalités

$$x^* = \lambda x, \quad y^* = \lambda y.$$

3) La contraction vers une droite (exemple 3 du § 1) s'écrit le plus aisément dans le repère cartésien rectangulaire dont l'axe des abscisses coïncide avec la droite vers laquelle est réalisée la contraction. On constate facilement que dans ce repère

$$x^* = x, \quad y^* = \lambda y,$$

où λ est le coefficient de contraction.

4) Le projecteur sur une droite (exemple 6 du § 1) s'écrit le plus aisément dans le repère cartésien rectangulaire dont l'axe des abscisses coïncide avec cette droite. On a donc dans ce repère,

$$x^* = x, \quad y^* = 0.$$

C'est une transformation linéaire mais non pas affine.

5) L'application associant à tout point du plan P un même point C du plan R s'écrit à l'aide des formules $x^* = c_1$, $y^* = c_2$, où c_1, c_2 sont les coordonnées du point C. Cette application est également linéaire mais non pas affine.

La définition de l'application linéaire présente le même défaut que celle de la courbe algébrique ; elle dépend du repère, et l'on ne sait pas si l'application linéaire sera définie par les formules (1) dans tout autre couple de repères cartésiens. Ce défaut est levé par la proposition suivante.

PROPOSITION 2. *Quels que soient les repères cartésiens dans les plans P et R, l'application linéaire $\mathbf{f} : P \to R$ est définie par les formules* (1).

DÉMONSTRATION. Supposons que l'application $\mathbf{f}$ est donnée par les égalités (1) dans les repères $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ et $\{Q, \mathbf{p}_1, \mathbf{p}_2\}$ des plans P et R respectivement. Passons aux nouveaux repères $\{O', \mathbf{e}_1', \mathbf{e}_2'\}$ et $\{Q', \mathbf{p}_1', \mathbf{p}_2'\}$. Exprimons les anciennes coordonnées d'un point du plan P au moyen des nouvelles par les formules

$$\left.\begin{aligned} x &= \alpha_1 x' + \beta_1 y' + \gamma_1, \\ y &= \alpha_2 x' + \beta_2 y' + \gamma_2, \end{aligned}\right\} \tag{3}$$

et dans le plan R, les nouvelles coordonnées au moyen des anciennes par les formules

$$\left.\begin{aligned} x^{*\prime} &= \lambda_1 x^* + \mu_1 y^* + \nu_1, \\ y^{*\prime} &= \lambda_2 x^* + \mu_2 y^* + \nu_2. \end{aligned}\right\} \tag{4}$$

Il nous faut exprimer les nouvelles coordonnées $x^{*\prime}$, $y^{*\prime}$ du point M^* par l'intermédiaire des nouvelles coordonnées x', y' du point M. A cet effet, portons dans (4) les expressions (1) pour les coordonnées de l'image M^* définies au moyen des coordonnées de son antécédent M. On obtient les nouvelles coordonnées de M^* exprimées en fonction des anciennes coordonnées de M :

$$\begin{aligned} x^{*\prime} &= \lambda_1(a_1 x + b_1 y + c_1) + \mu_1(a_2 x + b_2 y + c_2) + \nu_1, \\ y^{*\prime} &= \lambda_2(a_1 x + b_1 y + c_1) + \mu_2(a_2 x + b_2 y + c_2) + \nu_2. \end{aligned}$$

Il nous importe que les seconds membres de ces égalités sont des polynômes en x et y de degré inférieur ou égal à 1, c'est-à-dire sont de la forme

$$\left.\begin{aligned} x^{*\prime} &= A_1 x + B_1 y + C_1, \\ y^{*\prime} &= A_2 x + B_2 y + C_2. \end{aligned}\right\} \tag{5}$$

En portant dans (5) les expressions de x et y tirées des formules (3), on obtient la relation entre les nouvelles coordonnées de l'image et les nouvelles coordonnées de son antécédent, c'est-à-dire la dépendance cherchée

$$x^{*\prime} = A_1(\alpha_1 x' + \beta_1 y' + \gamma_1) + B_1(\alpha_2 x' + \beta_2 y' + \gamma_2) + C_1,$$
$$y^{*\prime} = A_2(\alpha_1 x' + \beta_1 y' + \gamma_1) + B_2(\alpha_2 x' + \beta_2 y' + \gamma_2) + C_2.$$

On voit que les seconds membres de ces égalités sont des polynômes en x' et y' de degré au plus égal à 1, c'est-à-dire que

$$x^{*\prime} = a_1' x' + b_1' y' + c_1',$$
$$y^{*\prime} = a_2' x' + b_2' y' + c_2'.$$

Ce qu'il fallait démontrer.

Signalons que la propriété des applications affines d'être bijectives ne dépend pas du choix des repères. Aussi la définition de l'application affine ne nécessite-t-elle pas d'arguments supplémentaires.

2. Produit d'applications linéaires. La démonstration de la proposition 2 s'est appuyée sur le fait que le résultat de la substitution de polynômes de degré au plus égal à 1 dans un polynôme de degré au plus égal à 1 est encore un polynôme de degré au plus égal à 1. Le même fait est à la base de la proposition suivante.

PROPOSITION 3. *Si le produit des applications linéaires est défini, il est une application linéaire.*

DÉMONSTRATION. Soient données les applications linéaires $\mathbf{f} : P \to R$ et $\mathbf{g} : R \to S$. Les plans P, R et S étant rapportés aux repères cartésiens, les coordonnées du point $\mathbf{f}(M)$ s'expriment en fonction des coordonnées du point M par les formules

$$\left.\begin{aligned} x^* &= a_1 x + b_1 y + c_1, \\ y^* &= a_2 x + b_2 y + c_2, \end{aligned}\right\} \tag{6}$$

et les coordonnées du point $\mathbf{g}(\mathbf{f}(M))$ se déterminent en fonction des coordonnées du point $\mathbf{f}(M)$ par les formules

$$\left.\begin{aligned} x^{**} &= d_1 x^* + e_1 y^* + f_1, \\ y^{**} &= d_2 x^* + e_2 y^* + f_2. \end{aligned}\right\} \tag{7}$$

La substitution de (6) dans (7) fournit l'expression des coordonnées de $\mathbf{g}(\mathbf{f}(M))$ au moyen des coordonnées de M :

$$x^{**} = d_1(a_1 x + b_1 y + c_1) + e_1(a_2 x + b_2 y + c_2) + f_1,$$
$$y^{**} = d_2(a_1 x + b_1 y + c_1) + e_2(a_2 x + b_2 y + c_2) + f_2.$$

On voit que les seconds membres sont des polynômes de degré inférieur ou égal à 1. Ce qui démontre notre assertion.

PROPOSITION 4. *Le produit des applications affines est une application affine.*

Il suffit, en vertu de la proposition 3, de démontrer que le produit des applications bijectives est aussi une application bijective. Or ce fait est assez évident. En effet, chaque point A du plan S a par l'application $\mathbf{g} : R \to S$ un antécédent B dans le plan R et un seul, qui, à son tour, admet par l'application $\mathbf{f} : P \to R$ un seul antécédent C dans le plan P. Le point C est justement l'antécédent unique du point A par l'application $\mathbf{g} \circ \mathbf{f}$. La proposition est démontrée.

Etant une application bijective, toute application affine admet une application réciproque.

PROPOSITION 5. *La réciproque de l'application affine est encore une application affine.*

Pour le démontrer, il nous faut résoudre l'équation (1) en x et y. Multiplions la première équation par b_2 et la seconde par $-b_1$ et additionnons-les. Il vient

$$(a_1b_2 - a_2b_1)x = b_2(x^* - c_1) - b_1(y^* - c_2).$$

Il ressort de la condition (2) que x est un polynôme linéaire en x^* et y^*. De façon analogue, on obtient l'expression pour y.

3. Image d'un vecteur par l'application linéaire. Considérons dans le plan P un vecteur $\overrightarrow{M_1M_2}$ et notons (x_1, y_1) et (x_2, y_2) les coordonnées des points M_1 et M_2 par rapport au repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$. Le vecteur $\overrightarrow{M_1M_2}$ a alors les composantes $(x_2 - x_1, y_2 - y_1)$. Supposons que les formules (1) définissent une application linéaire $\mathbf{f} : P \to R$ dans les repères $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ et $\{Q, \mathbf{p}_1, \mathbf{p}_2\}$. Dans ce cas, les images M_1^* et M_2^* des points M_1 et M_2 ont pour abscisses

$$x_1^* = a_1x_1 + b_1y_1 + c_1 \quad \text{et} \quad x_2^* = a_1x_2 + b_1y_2 + c_1.$$

Il s'ensuit que la première composante du vecteur $\overrightarrow{M_1^*M_2^*}$ est

$$x_2^* - x_1^* = a_1(x_2 - x_1) + b_1(y_2 - y_1).$$

De façon analogue, on obtient la seconde composante de $\overrightarrow{M_1^*M_2^*}$:

$$y_2^* - y_1^* = a_2(x_2 - x_1) + b_2(y_2 - y_1).$$

Soulignons la circonstance suivante : les composantes de $\overrightarrow{M_1^*M_2^*}$ ne s'expriment que par les composantes de $\overrightarrow{M_1M_2}$ et ne contiennent pas les coordonnées des points M_1 et M_2.

Considérons deux vecteurs égaux dans le plan P. Leurs composantes sont identiques et, par suite, l'application linéaire les transforme en vecteurs dont les composantes sont aussi identiques.

PROPOSITION 6. *Les vecteurs égaux se transforment par l'application linéaire en des vecteurs égaux. Les composantes* (α_1^*, α_2^*) *de l'image dans la base* $\{\mathbf{p}_1, \mathbf{p}_2\}$ *s'expriment alors en fonction des composantes* (α_1, α_2) *de l'antécédent dans la base* $\{\mathbf{e}_1, \mathbf{e}_2\}$ *par les formules*

$$\left.\begin{aligned} \alpha_1^* &= a_1\alpha_1 + b_1\alpha_2, \\ \alpha_2^* &= a_2\alpha_1 + b_2\alpha_2. \end{aligned}\right\} \tag{8}$$

Il découle de ces formules que l'application linéaire $\mathbf{f}$ vérifie les égalités

$$\left.\begin{aligned} \mathbf{f}(\mathbf{a} + \mathbf{b}) &= \mathbf{f}(\mathbf{a}) + \mathbf{f}(\mathbf{b}), \\ \mathbf{f}(\lambda\mathbf{a}) &= \lambda\mathbf{f}(\mathbf{a}), \end{aligned}\right\} \tag{9}$$

quels que soient les vecteurs $\mathbf{a}$ et $\mathbf{b}$ et le nombre α.

Démontrons d'abord la première des égalités (9). Soient γ_1^* et γ_2^* les composantes du vecteur $\mathbf{f}(\mathbf{a} + \mathbf{b})$. Il vient alors

$$\begin{aligned} \gamma_1^* &= a_1(\alpha_1 + \beta_1) + b_1(\alpha_2 + \beta_2), \\ \gamma_2^* &= a_2(\alpha_1 + \beta_1) + b_2(\alpha_2 + \beta_2), \end{aligned}$$

où (α_1, α_2) et (β_1, β_2) sont les composantes des vecteurs $\mathbf{a}$ et $\mathbf{b}$. D'où

$$\begin{aligned} \gamma_1^* &= a_1\alpha_1 + b_1\alpha_2 + a_1\beta_1 + b_1\beta_2 = \alpha_1^* + \beta_1^*, \\ \gamma_2^* &= a_2\alpha_1 + b_2\alpha_2 + a_2\beta_1 + b_2\beta_2 = \alpha_2^* + \beta_2^*. \end{aligned}$$

La proposition suivante révèle la signification géométrique des coefficients dans les formules (1).

PROPOSITION 7. *Soit une application* $\mathbf{f} : P \to R$ *définie par les formules* (1) *dans les repères* $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ et $\{Q, \mathbf{p}_1, \mathbf{p}_2\}$ *des plans* P *et* R. *Alors* (c_1, c_2) *sont les coordonnées du point* $\mathbf{f}(O)$ *par rapport au repère* $\{Q, \mathbf{p}_1, \mathbf{p}_2\}$, *tandis que* (a_1, a_2) *et* (b_1, b_2) *sont les composantes de* $\mathbf{f}(\mathbf{e}_1)$ *et* $\mathbf{f}(\mathbf{e}_2)$ *dans la base* $\{\mathbf{p}_1, \mathbf{p}_2\}$.

Portons dans (1) les valeurs $x = 0$ et $y = 0$, autrement dit les coordonnées du point O. On voit que les coordonnées de $\mathbf{f}(O)$ sont égales à c_1 et c_2.

Posons α_1 et α_2 de (8) égaux aux composantes de $\mathbf{e}_1$, autrement dit $\alpha_1 = 1, \alpha_2 = 0$. Alors $\alpha_1^* = a_1, \alpha_2^* = a_2$. Donc, $\mathbf{f}(\mathbf{e}_1)$ a pour composantes a_1, a_2. On démontre de façon analogue que les composantes de $\mathbf{f}(\mathbf{e}_2)$ sont égales à b_1 et b_2.

PROPOSITION 8. *Quels que soient trois points non alignés* L, M et N *du plan* P *et trois points* L^*, M^* *et* N^* *du plan* Q, *il existe une application linéaire* $\mathbf{f}$ *et une seule telle que* $L^* = \mathbf{f}(L)$, $M^* = \mathbf{f}(M)$ *et* $N^* = \mathbf{f}(N)$. *Cette application est affine si et seulement si* L^*, M^* *et* N^* *ne sont pas alignés.*

DÉMONSTRATION. Les vecteurs $\overrightarrow{LM}$ et $\overrightarrow{LN}$ ne sont pas colinéaires. Donc, $L, \overrightarrow{LM}, \overrightarrow{LN}$ forment un repère cartésien dans le plan P. Choisissons dans le plan R un repère arbitraire et supposons que c_1, c_2 sont les coordonnées de

L^*, et a_1, a_2 et b_1, b_2, les composantes des vecteurs $\overrightarrow{L^*M^*}$ et $\overrightarrow{L^*N^*}$ dans ce repère. Les formules

$$x^* = a_1x + b_1y + c_1,$$
$$y^* = a_2x + b_2y + c_2$$

définissent une application linéaire $\mathfrak{f} : P \to R$ qui, comme on le voit aisément, possède la propriété exigée : $L^* = \mathfrak{f}(L)$, $M^* = \mathfrak{f}(M)$, $N^* = \mathfrak{f}(N)$. De plus, en vertu de la proposition 7, cette propriété définit d'une façon univoque les coefficients intervenant dans les formules ci-dessus.

La condition (2) qui est une condition nécessaire et suffisante pour qu'une application linéaire soit affine est aussi nécessaire et suffisante pour que $\overrightarrow{L^*M^*}$ et $\overrightarrow{L^*N^*}$ ne soient pas colinéaires, autrement dit pour que L^*, M^* et N^* ne soient pas alignés.

PROPOSITION 9. *L'image M^* du point M par l'application affine $\mathfrak{f}$ a les mêmes coordonnées par rapport au repère $\{\mathfrak{f}(O), \mathfrak{f}(\mathbf{e}_1), \mathfrak{f}(\mathbf{e}_2)\}$ que le point M par rapport au repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$.*

Autrement dit, dans les repères $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ et $\{\mathfrak{f}(O), \mathfrak{f}(\mathbf{e}_1), \mathfrak{f}(\mathbf{e}_2)\}$ l'application affine se définit par les formules $x^* = x$, $y^* = y$. Cette assertion découle immédiatement de la proposition 7.

Choisissons dans les plans P et R deux repères cartésiens quelconques et associons à tout point M du plan P un point M^* du plan R de mêmes coordonnées. Cette relation est une application affine définie par les formules $x^* = x$, $y^* = y$. Il ressort de la proposition 9 que toute application affine peut être exprimée de cette manière.

§ 3. Transformations affines

1. Transformations orthogonales. Commençons l'étude des transformations affines par celles qui sont liées aux déplacements dans le plan. Ici et plus loin on les appellera *transformations orthogonales*. Démontrons d'abord que les trois principales transformations orthogonales, à savoir : la translation, la rotation et la symétrie axiale, sont des transformations affines.

a) La *translation* de vecteur $\mathbf{c}$ associe à tout point M de coordonnées x, y dans un repère cartésien donné un point M^* de coordonnées $x + c_1$, $y + c_2$, où c_1 et c_2 sont les composantes de $\mathbf{c}$. Il s'ensuit que

$$x^* = x + c_1, \quad y^* = y + c_2. \tag{1}$$

b) Considérons la *rotation* d'angle α du plan autour du point O. Dans le système de coordonnées polaires de pôle O, l'image M^* du point $M(r, \varphi)$ admet les coordonnées r et $\varphi + \alpha$. Les formules (3) du § 2, ch. I, nous per-

mettent de passer du système polaire au repère cartésien rectangulaire d'origine en O :

$$x^* = r\cos(\varphi + \alpha), \quad y^* = r\sin(\varphi + \alpha).$$

D'où, selon les formules du cosinus et du sinus de la somme de deux angles, il vient

$$\left.\begin{aligned} x^* &= x\cos\alpha - y\sin\alpha, \\ y^* &= x\sin\alpha + y\cos\alpha. \end{aligned}\right\} \tag{2}$$

c) Pour exprimer en coordonnées la *symétrie axiale*, choisissons un repère cartésien rectangulaire de façon que son axe des abscisses coïncide avec l'axe de symétrie. Tout point $M(x, y)$ se transforme alors en un point M^* de coordonnées x, $-y$. Ainsi donc, pour la symétrie axiale on a

$$x^* = x, \quad y^* = -y. \tag{3}$$

Les formules (1), (2) et (3) présentent, dans leurs seconds membres, des polynômes linéaires et définissent donc des transformations linéaires. Ces transformations sont bijectives, donc affines.

On définit une *transformation orthogonale* comme une transformation du plan conservant les distances entre les points. On sait qu'elle conserve également les angles entre les droites.

Il en découle que l'image par transformation orthogonale d'un repère cartésien rectangulaire $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ est un repère cartésien rectangulaire $\{O^*, \mathbf{e}_1^*, \mathbf{e}_2^*\}$, et que l'image d'un point M de coordonnées x, y par rapport au repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ est un point M^* de mêmes coordonnées x, y par rapport au repère $\{O^*, \mathbf{e}_1^*, \mathbf{e}_2^*\}$. Or les relations entre les coordonnées d'un même point par rapport à deux repères différents nous sont connues (formules (7), § 4, ch. I). Appliquées au point M^*, ces formules nous fournissent l'expression de ses coordonnées x^*, y^* dans le repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ en fonction de ses coordonnées x, y dans le repère $\{O^*, \mathbf{e}_1^*, \mathbf{e}_2^*\}$:

$$\left.\begin{aligned} x^* &= x\cos\varphi \mp y\sin\varphi + c_1, \\ y^* &= x\sin\varphi \pm y\cos\varphi + c_2. \end{aligned}\right\} \tag{4}$$

Ici (c_1, c_2) sont les coordonnées du point O^*, et φ est l'angle des vecteurs $\mathbf{e}_1$ et $\mathbf{e}_1^*$ mesuré dans le sens de $\mathbf{e}_1$ vers $\mathbf{e}_1^*$.

Rappelons que x, y sont les coordonnées de l'antécédent M. Cela signifie que les formules (4) peuvent être interprétées comme une expression analytique de la transformation orthogonale dans le repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$. On a ainsi démontré la

PROPOSITION 1. *Toute transformation orthogonale est affine et est définie dans tout repère cartésien rectangulaire à l'aide des formules* (4).

Cette proposition entraîne la

PROPOSITION 2. *Toute transformation orthogonale est le produit d'une rotation par une translation et, peut-être, encore par une symétrie axiale.*

En effet, soit une transformation orthogonale définie dans le repère cartésien rectangulaire par les formules (4). L'image par symétrie axiale d'un point $M(x, y)$ est un point $N(u, v)$, où $u = x$, $v = -y$ si l'axe de symétrie coïncide avec l'axe des abscisses de ce repère. La rotation d'angle φ autour de l'origine des coordonnées transforme le point N en un point $K(w, z)$, avec

$$w = u\cos\varphi - v\sin\varphi = x\cos\varphi + y\sin\varphi,$$
$$z = u\sin\varphi + v\cos\varphi = x\sin\varphi - y\cos\varphi.$$

Enfin, l'image du point $K(w, z)$ par translation de vecteur $\mathbf{c}(c_1, c_2)$ est un point $M^*(w + c_1, z + c_2)$ dont les coordonnées s'expriment en fonction de x, y par les formules (4), compte tenu des signes + et − inférieurs des coefficients affectant y. Si la symétrie axiale n'intervient pas, on obtient les formules (4) avec les signes + et − supérieurs.

On voit que le produit des transformations ainsi construit associe à un point arbitraire du plan la même image que la transformation orthogonale donnée. La proposition est démontrée.

Il faut remarquer que la représentation de la transformation orthogonale sous forme de produit n'est pas univoque. Par ailleurs, il est possible de décomposer une rotation ou une translation en produit de symétries axiales, et représenter le produit d'une translation par une rotation à l'aide d'une rotation unique, etc. Sans entrer dans ces détails proposons-nous de dégager la propriété générale suivante de ces décompositions.

PROPOSITION 3. *Quelle que soit la décomposition de la transformation orthogonale donnée en produit de rotations, translations et symétries axiales, la parité du nombre de symétries axiales figurant dans la décomposition est toujours la même.*

Pour le démontrer, considérons dans le plan une base quelconque et étudions la variation de son orientation (sens de la plus petite rotation de $\mathbf{e}_1$ vers $\mathbf{e}_2$) par transformation orthogonale. Signalons que la rotation et la translation ne modifient l'orientation d'aucune base, tandis que la symétrie axiale change l'orientation de toutes les bases. Il en découle que, si la transformation orthogonale donnée modifie l'orientation d'une base, toute décomposition de cette transformation doit comprendre un nombre impair de symétries axiales, et par suite, l'orientation de toute autre base doit être aussi modifiée. Mais si l'orientation de la base ne change pas, le nombre de symétries axiales dans la décomposition ne peut être que pair, et dans ce cas l'orientation de toute base demeure inchangée.

DÉFINITION. On dit qu'une transformation orthogonale est *de première espèce* si elle peut être décomposée en produit d'une rotation et d'une translation ; sinon, on dit qu'elle est *de deuxième espèce.*

PROPOSITION 4. *Les transformations orthogonales de première espèce se déterminent par les formules* (4) *avec des signes supérieurs des coefficients affectant y et ne modifient l'orientation d'aucune base. Les transformations orthogonales de deuxième espèce se déterminent par les formules* (4) *avec des signes inférieurs et modifient l'orientation de toute base.*

2. Image de la droite. Dans le texte qui suit, $\mathbf{f}$ désigne une transformation affine du plan, définie par rapport au repère cartésien $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ à l'aide des formules

$$\left.\begin{aligned} x^* &= a_1x + b_1y + c_1, \\ y^* &= a_2x + b_2y + c_2, \end{aligned}\right\} \tag{5}$$

à la condition que

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \neq 0. \tag{6}$$

Considérons dans le plan une droite d'équation $\mathbf{r} = \mathbf{r}_0 + \mathbf{a}t$ et cherchons son image par la transformation affine $\mathbf{f}$. Le rayon vecteur de l'image M^* d'un point arbitraire M de la droite peut être calculé de la façon suivante :

$$\mathbf{r}^* = \overrightarrow{OM^*} = \overrightarrow{O\mathbf{f}(O)} + \overrightarrow{\mathbf{f}(O)M^*} = \mathbf{c} + \mathbf{f}(\mathbf{r}),$$

où $\mathbf{c}$ est le vecteur constant $\overrightarrow{O\mathbf{f}(O)}$, et $\mathbf{r}$ le rayon vecteur du point M. En utilisant les propriétés (9), données au § 2, des transformations affines, on obtient

$$\mathbf{r}^* = \mathbf{c} + \mathbf{f}(\mathbf{r}_0) + \mathbf{f}(\mathbf{a})t. \tag{7}$$

L'image par transformation affine d'un vecteur non nul est un vecteur non nul, car autrement il existerait deux points différents ayant une même image, ce qui est impossible pour une transformation bijective. Par conséquent, $\mathbf{f}(\mathbf{a}) \neq \mathbf{0}$ et l'équation (7) est une équation de la droite. Ainsi donc, les images de tous les points de la droite $\mathbf{r} = \mathbf{r}_0 + \mathbf{a}t$ se trouvent sur la droite (7).

Notons que la transformation $\mathbf{f}$ engendre une bijection entre les points des deux droites. Etant donné un point initial et un vecteur directeur sur chacune d'elles, l'image du point M admet sur la droite (7) la même valeur du paramètre t que le point M sur la droite initiale. D'où la

PROPOSITION 5. *L'image par transformation affine d'une droite est une droite, celle d'un segment est un segment, et les images de deux droites parallèles sont deux droites parallèles.*

Pour démontrer que tout segment se transforme en un segment, signalons que le segment de droite est composé de points dont les valeurs du paramètre t vérifient la double inégalité $t_1 \leqslant t \leqslant t_2$. La troisième assertion de la proposition découle du fait qu'en vertu des formules (9), § 2, les vecteurs colinéaires se transforment en des vecteurs colinéaires.

PROPOSITION 6. *La transformation affine ne change pas le rapport des longueurs de deux segments parallèles.*

DÉMONSTRATION. Soient deux segments parallèles AB et CD. Cela signifie qu'il existe un nombre λ tel que $\overrightarrow{AB} = \lambda\overrightarrow{CD}$, de sorte que les images des vecteurs $\overrightarrow{AB}$ et $\overrightarrow{CD}$ sont liées par la même relation $\overrightarrow{A^*B^*} = \lambda\overrightarrow{C^*D^*}$. Il en découle que

$$\frac{|AB|}{|CD|} = \frac{|A^*B^*|}{|C^*D^*|} = |\lambda|.$$

COROLLAIRE. *Si le point C partage le segment AB dans un rapport λ, son image C^* partage l'image A^*B^* du segment AB dans le même rapport λ.*

3. Variation de l'aire d'une figure par transformation affine. Soit pour commencer un parallélogramme quelconque construit sur les vecteurs $\mathbf{p}$ et $\mathbf{q}$. Choisissons un repère cartésien $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ et désignons par (x_1, y_1) et (x_2, y_2) les composantes de $\mathbf{p}$ et $\mathbf{q}$. On peut calculer l'aire du parallélogramme en se servant des propriétés du produit vectoriel :

$$S = |[\mathbf{p}, \mathbf{q}]| = |[x_1\mathbf{e}_1 + y_1\mathbf{e}_2, x_2\mathbf{e}_1 + y_2\mathbf{e}_2]| = |x_1y_2 - y_1x_2|\,|[\mathbf{e}_1, \mathbf{e}_2]|.$$

Admettons que la transformation affine $\mathbf{f}$ rapportée au repère choisi est définie par les formules (5). Il ressort de la proposition 7 du § 2 que $\mathbf{f}(\mathbf{e}_1) = a_1\mathbf{e}_1 + a_2\mathbf{e}_2$ et $\mathbf{f}(\mathbf{e}_2) = b_1\mathbf{e}_1 + b_2\mathbf{e}_2$. Selon la proposition 9 du § 2, les vecteurs $\mathbf{f}(\mathbf{p})$ et $\mathbf{f}(\mathbf{q})$ rapportés à la base $\{\mathbf{f}(\mathbf{e}_1), \mathbf{f}(\mathbf{e}_2)\}$ possèdent les mêmes composantes (x_1, y_1) et (x_2, y_2) que les vecteurs $\mathbf{p}$ et $\mathbf{q}$ rapportés à la base $\{\mathbf{e}_1, \mathbf{e}_2\}$. L'image du parallélogramme est construite sur les vecteurs $\mathbf{f}(\mathbf{p})$ et $\mathbf{f}(\mathbf{q})$ et son aire est

$$\begin{aligned} S^* = |[\mathbf{f}(\mathbf{p}), \mathbf{f}(\mathbf{q})]| &= |x_1y_2 - x_2y_1|\,|[\mathbf{f}(\mathbf{e}_1), \mathbf{f}(\mathbf{e}_2)]| = \\ &= |x_1y_2 - x_2y_1|\,|[a_1\mathbf{e}_1 + a_2\mathbf{e}_2, b_1\mathbf{e}_1 + b_2\mathbf{e}_2]| = \\ &= |x_1y_2 - x_2y_1|\,|a_1b_2 - a_2b_1|\,|[\mathbf{e}_1, \mathbf{e}_2]|. \end{aligned}$$

D'où en définitive,

$$S^*/S = |a_1b_2 - a_2b_1|. \tag{8}$$

Cette expression montre que le rapport des aires S^* et S est le même pour tous les parallélogrammes et ne dépend pas du repère dans lequel la transformation est définie, bien que dans son expression $|a_1b_2 - a_2b_1|$ figurent

les coefficients dépendant du repère. Ce rapport est un invariant (voir p. 49) traduisant la propriété géométrique de la transformation.

Soit $x_1y_2 - x_2y_1 > 0$. Alors, comme on l'a vu au point 8 du § 3, ch. I, les couples des vecteurs **p**, **q** et $\mathbf{e}_1$, $\mathbf{e}_2$ ont une même orientation, et il en est de même des couples $\mathbf{f}(\mathbf{p})$, $\mathbf{f}(\mathbf{q})$ et $\mathbf{f}(\mathbf{e}_1)$, $\mathbf{f}(\mathbf{e}_2)$. Par conséquent, si la transformation affine change l'orientation des vecteurs de base, elle change celle du couple **p**, **q**. On démontre de façon analogue que la transformation affine change l'orientation de tout couple dont l'orientation est opposée à celle du couple des vecteurs de base.

Mais si l'orientation de la base reste inchangée par transformation affine, il en est de même pour les autres couples ordonnés de vecteurs.

Cette propriété ne dépend évidemment pas du choix du couple de vecteurs de base : ou bien l'orientation change pour tous les couples de vecteurs, ou bien elle ne change pour aucun d'eux. On a déjà démontré cette propriété pour les transformations orthogonales.

Signalons maintenant que le couple des vecteurs $\mathbf{f}(\mathbf{e}_1)$, $\mathbf{f}(\mathbf{e}_2)$ est orienté de la même façon que $\mathbf{e}_1$, $\mathbf{e}_2$ si et seulement si $a_1b_2 - a_2b_1 > 0$. On en conclut que non seulement le module du déterminant $a_1b_2 - a_2b_1$ mais aussi son signe ne dépendent pas du choix du repère. On peut donc affirmer que $a_1b_2 - a_2b_1$ est un invariant lié à la transformation affine et dont la signification géométrique est le rapport des aires de deux parallélogrammes orientés, où le premier parallélogramme est l'image du second par transformation affine étudiée.

Passons maintenant aux aires d'autres figures. Tout triangle peut être complété jusqu'à un parallélogramme dont l'aire est le double de celle du triangle. Il en ressort que le rapport de l'aire de l'image du triangle à celle du triangle donné vérifie l'égalité (8).

Tout polygone peut être subdivisé en triangles. Donc, la formule (8) est également vraie pour les aires de polygones quelconques.

On ne touchera pas ici à la définition de l'aire des figures curvilignes. Indiquons seulement que dans les cas où cette aire est définie, elle est égale à la limite des aires d'une suite de polygones inscrits dans la figure considérée. La théorie de la limite dit que : si une suite de S_n converge vers S, la suite de δS_n, où δ est une constante, converge vers δS. En s'appuyant sur cette proposition, on peut conclure que la formule (8) reste vraie dans le cas le plus général.

A titre d'exemple, cherchons l'aire de l'ellipse en fonction de ses demi-axes. On a démontré dans le § 2 du ch. II que l'ellipse de demi-axes a et b peut être obtenue par contraction du cercle de rayon a vers une droite passant par son centre. Le coefficient de contraction est b/a. Ceci étant, un carré de côté 1 et dont la base est située sur la droite mentionnée se transforme évidemment en un rectangle de côtés b/a et 1. Donc, le rapport de l'aire de l'image du carré à l'aire de ce carré vaut b/a. Or on a vu que la transformation affine modifie les aires de toutes les figures dans le même rapport. Aussi a-t-on pour l'aire de l'ellipse

$$S = (\pi a^2)(b/a).$$

D'où il vient

$$S = \pi ab.$$

4. Images des coniques. On a vu qu'une droite se transformait en une droite. Ce fait peut être généralisé par la proposition suivante.

PROPOSITION 7. *L'image par transformation affine d'une courbe algébrique est une courbe algébrique de même ordre.*

En effet, soit la courbe L définie dans le repère cartésien $\{O, \mathbf{e}_1, \mathbf{e}_2\}$ par une équation algébrique de degré p. Selon la proposition 9 du § 2, les images par la transformation affine $\mathbf{f}$ de tous les points de la courbe L possèdent dans le repère $\{\mathbf{f}(O), \mathbf{f}(\mathbf{e}_1), \mathbf{f}(\mathbf{e}_2)\}$ les mêmes coordonnées que leurs antécédents dans le repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$. Par conséquent, les coordonnées des images dans le repère $\{\mathbf{f}(O), \mathbf{f}(\mathbf{e}_1), \mathbf{f}(\mathbf{e}_2)\}$ sont liées par la même équation algébrique de degré p. Cela signifie que l'image de la courbe L est définie dans le repère cartésien $\{\mathbf{f}(O), \mathbf{f}(\mathbf{e}_1), \mathbf{f}(\mathbf{e}_2)\}$ par l'équation algébrique de degré p. L'image de la courbe L est donc une courbe algébrique de même ordre que la courbe L.

Il résulte, en particulier, de la proposition 7 que l'image par transformation affine de toute conique est une conique. On démontrera une assertion plus forte, à savoir : toute transformation affine laisse invariante chacune des sept classes en lesquelles, selon le théorème 1 du § 1, ch. III, on a réparti les coniques. C'est pourquoi, les classes des coniques du théorème sont dites *affines*. Plus précisément, on a la

PROPOSITION 8. *L'image par toute transformation affine de chaque conique d'une classe affine donnée est encore une conique de cette classe. Par ailleurs, à chaque conique on peut associer, par une transformation affine appropriée, toute autre conique de la même classe.*

DÉMONSTRATION. On dit que la conique est bornée s'il existe un parallélogramme qui la contient. On constate aisément que dans une transformation affine, l'image de toute conique bornée est une conique bornée et celle d'une conique non bornée est non bornée.

1) L'ellipse est une conique bornée. Outre les ellipses ne sont bornées que les coniques composées d'un seul point et appelées couples de droites sécantes imaginaires. Puisque l'ellipse est bornée et se compose de plus d'un point, son image par toute transformation affine est une ellipse.

2) L'hyperbole comprend deux branches séparées. Cette propriété peut être formulée de manière que son invariabilité par transformation affine devienne évidente. Plus précisément, il existe une droite qui ne coupe pas l'hyperbole mais coupe plusieurs de ses cordes (c'est-à-dire que les points de l'hyperbole se trouvent de part et d'autre de cette droite).

De toutes les coniques, seules les hyperboles et les couples de droites parallèles sont composées de deux branches. Les branches de l'hyperbole ne sont pas des droites et, par suite, l'image par transformation affine d'une hyperbole n'est qu'une hyperbole.

3) La parabole est une conique non bornée composée d'une seule branche curviligne. Aucune autre conique ne possède cette propriété, aussi l'image par transformation affine d'une parabole n'est-elle qu'une parabole.

4) Si une conique représente un point (couple de droites sécantes imaginaires), une droite (couple de droites confondues), un couple de droites sécantes ou parallèles, les propriétés déjà démontrées des transformations affines entraînent que son image ne peut appartenir à aucune autre classe affine.

Démontrons maintenant que pour tout couple de coniques d'une même classe affine il existe une transformation affine dans laquelle l'une des coniques est l'image de l'autre. Pour le faire, remarquons que dans le théorème 1 du § 1, ch. III, les équations canoniques sont rapportées à un repère cartésien rectangulaire et contiennent les paramètres a, b, ... Donc, chaque équation citée dans le théorème représente en fait un ensemble d'équations correspondant à différentes valeurs des paramètres. Si l'on cesse d'exiger que la base soit orthonormée, tout cet ensemble peut être réduit à la même forme canonique démunie de paramètres. Par exemple, le changement de coordonnées $x' = x/a$, $y' = y/b$ transforme l'équation de l'ellipse $x^2/a^2 + y^2/b^2 = 1$ en $x'^2 + y'^2 = 1$ quels que soient a et b. (La dernière équation n'est pas celle d'un cercle car le repère cartésien n'est plus rectangulaire.) Le lecteur montrera aisément que pour toute équation du second degré il existe un repère cartésien dans lequel cette équation prend l'une des formes suivantes :

1) $x^2 + y^2 = 1$, 2) $x^2 + y^2 = -1$, 3) $x^2 - y^2 = 1$,
4) $x^2 - y^2 = 0$, 5) $x^2 + y^2 = 0$, 6) $y^2 = 2x$,
7) $y^2 - 1 = 0$, 8) $y^2 + 1 = 0$, 9) $y^2 = 0$.

Un tel repère sera appelé *repère canonique affine*.

Il ressort de la proposition 9 du § 2 que si une transformation affine fait coïncider les repères canoniques affines de deux courbes d'une même classe, il en est de même de ces courbes. La proposition est ainsi complètement démontrée.

PROPOSITION 9. *Pour tout couple de coniques d'une même classe affine, dont les équations canoniques présentent les mêmes valeurs de paramètres, il existe une transformation orthogonale qui les fait coïncider.*

En effet, la transformation orthogonale les faisant coïncider est celle qui confond leurs repères canoniques rectangulaires.

5. Description de toutes les transformations affines. On a vu à quel point une transformation affine peut modifier toutes les figures : un cercle peut passer à une ellipse, un triangle équilatérale à un triangle quelconque.

Il semble qu'aucun angle n'est alors invariant. Or il s'avère de façon inattendue qu'on a la

PROPOSITION 10. *Dans toute transformation affine il existe deux droites perpendiculaires auxquelles correspondent deux droites perpendiculaires.*

Pour le démontrer, considérons un cercle quelconque. Il passe par une transformation affine étudiée à une ellipse. Chaque axe de symétrie de l'ellipse est un ensemble des milieux des cordes parallèles à l'autre axe. On sait que dans une transformation affine l'image d'une corde est encore une corde, que les images de deux cordes parallèles sont aussi parallèles et que le milieu de la corde redevient son milieu (proposition 5). Aussi les antécédents des axes de symétrie de l'ellipse sont-ils des segments ayant la même propriété : chacun d'eux est un ensemble des milieux des cordes du cercle qui sont parallèles à l'autre segment. Ces segments sont nécessairement deux diamètres perpendiculaires du cercle. Ainsi donc, la proposition est démontrée : il existe deux diamètres perpendiculaires du cercle dont les images sont deux segments perpendiculaires : les axes de symétrie de l'ellipse.

Cette proposition nous permet de décrire toutes les transformations affines.

THÉORÈME 1. *Toute transformation affine est le produit d'une transformation orthogonale et de deux contractions vers les droites perpendiculaires.*

DÉMONSTRATION. Notons $\mathbf{e}_1$ et $\mathbf{e}_2$ deux vecteurs perpendiculaires dont les images par une transformation affine $\mathbf{f}$ sont des vecteurs perpendiculaires. Ces vecteurs existent, selon la proposition 10. Considérons un point arbitraire O et le repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$. L'image de ce repère par $\mathbf{f}$ est le repère $\{\mathbf{f}(O), \mathbf{f}(\mathbf{e}_1), \mathbf{f}(\mathbf{e}_2)\}$ (fig. 51). A l'aide d'une transformation orthogonale $\mathbf{g}$ on peut associer au point O le point $\mathbf{f}(O)$, ainsi que faire coïncider les direc-

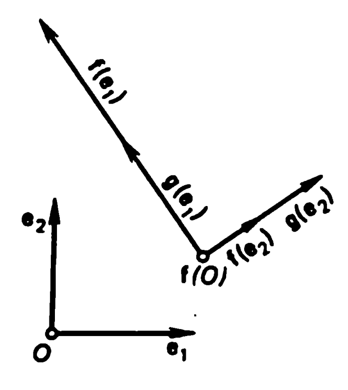


Fig. 51.

tions des vecteurs de base $\mathbf{e}_1$ et $\mathbf{e}_2$ avec les directions de leurs images, c'est-à-dire choisir $\mathbf{g}$ de manière que

$$\mathbf{g}(O) = \mathbf{f}(O), \quad \lambda_1 \mathbf{g}(\mathbf{e}_1) = \mathbf{f}(\mathbf{e}_1), \quad \lambda_2 \mathbf{g}(\mathbf{e}_2) = \mathbf{f}(\mathbf{e}_2),$$
$$\lambda_1, \lambda_2 > 0.$$

Soit $\mathbf{h}_1$ la contraction dans le rapport λ_1 vers la droite passant par $\mathbf{f}(O)$ en direction du vecteur $\mathbf{f}(\mathbf{e}_2)$. Cette contraction transforme $\mathbf{g}(\mathbf{e}_1)$ en $\mathbf{f}(\mathbf{e}_1)$ sans changer ni le point $\mathbf{f}(O)$ ni le vecteur $\mathbf{g}(\mathbf{e}_2)$:

$$\mathbf{h}_1 \circ \mathbf{g}(\mathbf{e}_1) = \mathbf{f}(\mathbf{e}_1), \quad \mathbf{h}_1 \circ \mathbf{g}(O) = \mathbf{f}(O), \quad \mathbf{h}_1 \circ \mathbf{g}(\mathbf{e}_2) = \mathbf{g}(\mathbf{e}_2).$$

Soit $\mathbf{h}_2$ la contraction dans le rapport λ_2 vers la droite passant par $\mathbf{f}(O)$ en direction du vecteur $\mathbf{f}(\mathbf{e}_1)$. Elle transforme $\mathbf{g}(\mathbf{e}_2)$ en $\mathbf{f}(\mathbf{e}_2)$ et ne modifie pas le point $\mathbf{f}(O)$ et le vecteur $\mathbf{f}(\mathbf{e}_1)$. Donc

$$\mathbf{h}_2 \circ \mathbf{h}_1 \circ \mathbf{g}(O) = \mathbf{f}(O), \quad \mathbf{h}_2 \circ \mathbf{h}_1 \circ \mathbf{g}(\mathbf{e}_1) = \mathbf{f}(\mathbf{e}_1),$$
$$\mathbf{h}_2 \circ \mathbf{h}_1 \circ \mathbf{g}(\mathbf{e}_2) = \mathbf{f}(\mathbf{e}_2).$$

On voit que les transformations $\mathbf{h}_2 \circ \mathbf{h}_1 \circ \mathbf{g}$ et $\mathbf{f}$ associent le même repère au repère $\{O, \mathbf{e}_1, \mathbf{e}_2\}$. On sait de la proposition 8 du § 2 qu'une transformation affine est entièrement définie par l'image d'un repère cartésien. Aussi les transformations affines $\mathbf{f}$ et $\mathbf{h}_2 \circ \mathbf{h}_1 \circ \mathbf{g}$ coïncident-elles et le théorème est démontré.

CHAPITRE V

SYSTÈMES D'ÉQUATIONS LINÉAIRES ET MATRICES

§ 1. Matrices

1. Définition. On appelle *matrice à m lignes et n colonnes*, ou *matrice de type* (m, n), un ensemble de mn nombres rangés en tableau à m lignes et n colonnes :

$$\left\|\begin{matrix} a_1^1 & a_2^1 & \dots & a_n^1 \\ a_1^2 & a_2^2 & \dots & a_n^2 \\ \dots & \dots & \dots & \dots \\ a_1^m & a_2^m & \dots & a_n^m \end{matrix}\right\|.$$

Les nombres intervenant dans la matrice sont appelés *éléments* de la matrice. Si le nombre des lignes est égal au nombre des colonnes, la matrice est dite *carrée*, et le nombre des lignes est appelé *ordre* de la matrice. Les autres matrices sont dénommées *rectangulaires*.

Les matrices peuvent aussi être définies de la façon suivante.

Considérons deux ensembles de nombres entiers : $I = \{1, 2, \dots, m\}$ et $J = \{1, 2, \dots, n\}$. Notons $I \times J$ l'ensemble de tous les couples (i, j), où i est un nombre de I et j un nombre de J. On appelle *matrice* une application définie sur l'ensemble $I \times J$, c'est-à-dire une relation associant à chaque couple (i, j) un nombre a_j^i.

Deux matrices sont dites *égales* si elles ont mêmes dimensions et si sont égaux leurs éléments homologues.

En étudiant les matrices, on désignera leurs éléments par des lettres affectées de deux indices. Si les deux indices sont inférieurs, le premier désigne la ligne, le second, la colonne ; si l'un des indices est supérieur, comme c'est le cas de la matrice mentionnée ci-dessus, il indique le numéro de la ligne. Attention : ne pas confondre cet indice avec un exposant.

Il est souvent commode d'envisager la matrice comme un ensemble de colonnes ou un ensemble de lignes. Soient

$$a_1 = \left\|\begin{matrix} a_1^1 \\ a_1^2 \\ \dots \\ a_1^m \end{matrix}\right\|, \quad a_2 = \left\|\begin{matrix} a_2^1 \\ a_2^2 \\ \dots \\ a_2^m \end{matrix}\right\|, \quad \dots, \quad a_n = \left\|\begin{matrix} a_n^1 \\ a_n^2 \\ \dots \\ a_n^m \end{matrix}\right\|.$$

Alors la matrice écrite initialement peut être présentée sous la forme

$$\| a_1 a_2 \dots a_n \| .$$

De façon analogue, si

$$a^1 = \| a_1^1 \dots a_n^1 \|, \dots, a^m = \| a_1^m \dots a_n^m \|,$$

la même matrice s'écrit sous la forme

$$\left\| \begin{matrix} a^1 \\ \dots \\ a^m \end{matrix} \right\| .$$

2. Addition et multiplication par un nombre. Soient A et B deux matrices à m lignes et n colonnes. On peut leur associer une troisième matrice C à m lignes et n colonnes, dont chaque élément est la somme des éléments de mêmes indices des matrices A et B. Autrement dit, les éléments c_{ij} de la matrice C sont liés aux éléments a_{ij} et b_{ij} des matrices A et B par l'égalité

$$c_{ij} = a_{ij} + b_{ij} \tag{1}$$

pour tous $i = 1, \dots, m$ et $j = 1, \dots, n$.

DÉFINITION. La matrice C définie pour A et B par la formule (1) est appelée leur *somme* et est notée $A + B$.

Soulignons que la somme n'est définie que pour des matrices de mêmes type.

La définition de la somme des matrices correspond entièrement à celle de la somme des fonctions : sont additionnées des valeurs correspondant à un même élément de l'ensemble de définition, c'est-à-dire à un même couple (i, j).

DÉFINITION. La matrice C dont les éléments c_{ij} sont égaux aux produits des éléments a_{ij} de la matrice A par le nombre α s'appelle *produit* de A par α et se note αA. On a

$$c_{ij} = \alpha a_{ij} \tag{2}$$

pour tous $i = 1, \dots, m$; $j = 1, \dots, n$.

Des propriétés d'addition et de multiplication des nombres on déduit aisément la proposition suivante.

PROPOSITION 1. *Pour toutes matrices A, B et C de même type et tous nombres α et β sont vérifiées les égalités*

$$A + B = B + A, \quad (A + B) + C = A + (B + C),$$
$$\alpha(A + B) = \alpha A + \alpha B, \quad (\alpha + \beta)A = \alpha A + \beta A,$$
$$(\alpha\beta)A = \alpha(\beta A).$$

La matrice dont tous les éléments sont nuls est dite *nulle*. Si O est la matrice nulle à m lignes et n colonnes, on a pour toute matrice A de type (m, n)

$$A + O = A.$$

La matrice $(-1)A$ sera appelée *matrice opposée* à A et notée $-A$. Elle possède la propriété suivante :

$$A + (-A) = O.$$

La somme des matrices B et $-A$ s'appelle *différence* des matrices B et A. On la notera $B - A$.

3. Transposition des matrices. Soit la matrice

$$A = \begin{Vmatrix} a_{11} & \dots & a_{1n} \\ a_{21} & \dots & a_{2n} \\ \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} \end{Vmatrix}$$

à m lignes et n colonnes. On peut lui associer la matrice B à n lignes et m colonnes en se conformant à la règle suivante. On écrit les éléments de chaque ligne de la matrice A dans la colonne de la matrice B sans changer leur ordre, le numéro de la colonne coïncidant avec celui de la ligne. Il est évident que dans ce cas la i-ième ligne de B est composée des mêmes éléments pris dans le même ordre que la i-ième colonne de A. La matrice

$$B = \begin{Vmatrix} a_{11} & a_{21} & \dots & a_{m1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{Vmatrix}$$

est dite la *transposée* de A et notée ${}^tA$. Le passage de A à ${}^tA$ est appelé *transposition*.

La définition d'une matrice transposée peut être écrite sous la forme de mn égalités

$$b_{ji} = a_{ij},$$

reliant les éléments des matrices A et B pour tous les $i = 1, \dots, m$ et $j = 1, \dots, n$.

4. Matrices-colonnes et matrices-lignes. La matrice de type $(1, n)$, c'est-à-dire la matrice composée d'une seule ligne sera appelée *matrice-ligne* à n éléments. La matrice de type $(m, 1)$ composée d'une seule colonne sera appelée *matrice-colonne* à m éléments. L'addition des matrices-lignes n'est définie que pour des matrices-lignes à même nombre d'éléments. Il en

est de même de l'addition des matrices-colonnes, qui n'est définie que pour des matrices-colonnes à même nombre d'éléments. Pour ces matrices, l'addition et la multiplication par un nombre seront étudiées d'une manière plus détaillée. Vu que toutes les propriétés des matrices-lignes et des matrices-colonnes sont formulées et démontrées d'une façon analogue, seul sera traité le cas des matrices-colonnes.

Les matrices-colonnes et les matrices-lignes seront notées en caractères demi-gras.

DÉFINITION. La matrice-colonne $\boldsymbol{q}$ est appelée *combinaison linéaire des matrices-colonnes* $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m$ ayant un même nombre d'éléments s'il existe des nombres $\alpha_1, \ldots, \alpha_m$ tels que l'on ait

$$\boldsymbol{q} = \sum_{k=1}^{m} \alpha_k \boldsymbol{p}_k$$

ou d'une façon plus détaillée

$$\left\| \begin{matrix} q^1 \\ q^2 \\ \ldots \\ q^n \end{matrix} \right\| = \alpha_1 \left\| \begin{matrix} p_1^1 \\ p_1^2 \\ \ldots \\ p_1^n \end{matrix} \right\| + \alpha_2 \left\| \begin{matrix} p_2^1 \\ p_2^2 \\ \ldots \\ p_2^n \end{matrix} \right\| + \ldots + \alpha_m \left\| \begin{matrix} p_m^1 \\ p_m^2 \\ \ldots \\ p_m^n \end{matrix} \right\| .$$

Signalons qu'en vertu des définitions de l'addition et de la multiplication par un nombre, cette égalité est équivalente à n égalités numériques

$$\begin{aligned} q^1 &= \alpha_1 p_1^1 + \alpha_2 p_2^1 + \ldots + \alpha_m p_m^1, \\ q^2 &= \alpha_1 p_1^2 + \alpha_2 p_2^2 + \ldots + \alpha_m p_m^2, \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ q^n &= \alpha_1 p_1^n + \alpha_2 p_2^n + \ldots + \alpha_m p_m^n. \end{aligned}$$

On a déjà utilisé l'expression « combinaison linéaire » pour les vecteurs. La ressemblance de deux définitions n'est pas seulement formelle. La base étant choisie, à tout vecteur on peut faire correspondre une matrice-ligne formée des composantes (au nombre de trois) de ce vecteur, et à toute combinaison linéaire des vecteurs une combinaison linéaire des matrices-lignes formées des composantes de ces vecteurs. On poursuivra l'analogie entre les vecteurs et les matrices-lignes (resp. matrices-colonnes) après avoir défini la notion de dépendance linéaire. Désignons par le symbole $\boldsymbol{o}$ la matrice-colonne dont tous les éléments sont nuls.

DÉFINITION. Le système de s matrices-colonnes $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_s$ de même type est dit *linéairement indépendant* ou *libre* si l'égalité

$$\alpha_1 \boldsymbol{a}_1 + \ldots + \alpha_s \boldsymbol{a}_s = \boldsymbol{o} \tag{3}$$

entraîne $\alpha_1 = \alpha_2 = \ldots = \alpha_s = 0$. Dans le cas contraire, c'est-à-dire s'il existe s nombres $\alpha_1, \ldots, \alpha_s$ simultanément non nuls et vérifiant l'égalité (3), le système $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_s$ est dit ***linéairement dépendant*** ou ***lié***.

Les définitions d'un système de matrices-lignes linéairement dépendant et linéairement indépendant sont formulées de façon identique.

La combinaison linéaire dont tous les coefficients sont nuls est dite *triviale*. Ce terme nous permet de formuler la définition précédente comme suit : le système de matrices-colonnes est linéairement dépendant ou lié s'il existe une combinaison linéaire non triviale de ces matrices-colonnes qui est égale à zéro. Le système de matrices-colonnes est linéairement indépendant ou libre si la seule combinaison linéaire triviale de ces matrices-colonnes est égale à zéro.

EXEMPLE. Les matrices-colonnes

$$\boldsymbol{e}_1 = \begin{Vmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{Vmatrix}, \quad \boldsymbol{e}_2 = \begin{Vmatrix} 0 \\ 1 \\ \ldots \\ 0 \end{Vmatrix}, \quad \ldots, \quad \boldsymbol{e}_n = \begin{Vmatrix} 0 \\ \ldots \\ 0 \\ 1 \end{Vmatrix} \tag{4}$$

(dans la matrice-colonne $\boldsymbol{e}_i$, le i-ième élément est 1, les autres éléments étant nuls) sont linéairement indépendantes. En effet, l'égalité $\alpha_1 \boldsymbol{e}_1 + \ldots$ $\ldots + \alpha_n \boldsymbol{e}_n = \boldsymbol{o}$ peut être écrite sous la forme

$$\alpha_1 \begin{Vmatrix} 1 \\ 0 \\ \ldots \\ 0 \end{Vmatrix} + \ldots + \alpha_n \begin{Vmatrix} 0 \\ \ldots \\ 0 \\ 1 \end{Vmatrix} = \begin{Vmatrix} \alpha_1 \\ \ldots \\ \alpha_n \end{Vmatrix} = \begin{Vmatrix} 0 \\ \ldots \\ 0 \end{Vmatrix}.$$

Il s'ensuit que $\alpha_1 = \ldots = \alpha_n = 0$.

Mentionnons quelques propriétés des systèmes liés de matrices-colonnes. Les propositions 2, 3, 4 ci-après sont analogues aux propositions qui ont été formulées au chapitre I pour les vecteurs, leurs démonstrations coïncident avec les démonstrations données dans ce chapitre.

PROPOSITION 2. *Un système de $s > 1$ matrices-colonnes est lié si et seulement si l'une au moins des ces matrices-colonnes est une combinaison linéaire des autres.*

En effet, supposons que le système est lié. En vertu de la définition, l'égalité (3) est vérifiée et l'un au moins de ses coefficients est différent de zéro. Posons, pour fixer les idées, que c'est α_1. On peut alors écrire l'égalité (3) sous la forme

$$\boldsymbol{a}_1 = -\frac{\alpha_2}{\alpha_1}\boldsymbol{a}_2 - \ldots - \frac{\alpha_s}{\alpha_1}\boldsymbol{a}_s.$$

On en tire que la première matrice-colonne est une combinaison linéaire des autres.

Inversement, quand une des matrices-colonnes (soit a_1 pour fixer les idées) est une combinaison linéaire des autres, on a $a_1 = \beta_2 a_2 + \dots + \beta_s a_s$. Si l'on porte tous les termes de cette égalité dans l'un de ses membres, on pourra traiter cette relation comme une combinaison linéaire non triviale des matrices-colonnes $a_1, \dots, a_s$ qui est égale à zéro.

Il découle de la démonstration que chaque matrice-colonne intervenant avec un coefficient non nul dans la combinaison linéaire qui est égale à zéro peut être représentée comme une combinaison linéaire des autres matrices-colonnes.

PROPOSITION 3. *Le système de matrices-colonnes est lié s'il contient la matrice-colonne nulle.*

En effet, la matrice-colonne nulle représente une combinaison linéaire triviale de matrices-colonnes quelconques, de sorte que la démonstration se réduit à la proposition 2.

PROPOSITION 4. *Le système des matrices-colonnes $a_1, \dots, a_s$ est lié si l'un quelconque de ses sous-systèmes est lié.*

Admettons que le système $\{a_1, \dots, a_s\}$ contient des matrices-colonnes dont une combinaison linéaire non triviale est égale à zéro. Si on ajoute à cette dernière les autres matrices-colonnes avec coefficients nuls, on obtient une combinaison linéaire non triviale égale à zéro de toutes les matrices-colonnes.

PROPOSITION 5. *Chaque sous-système d'un système de matrices-colonnes libre est libre.*

En effet, dans le cas contraire, on aboutirait à une affirmation qui contredit la proposition précédente.

PROPOSITION 6. *Si la matrice colonne a est une combinaison linéaire des matrices-colonnes $a_1, \dots, a_s$, elle est aussi une combinaison linéaire de tout système de matrices-colonnes qui contient $a_1, \dots, a_s$.*

Pour le démontrer, il suffit d'ajouter à la combinaison linéaire donnée les matrices-colonnes manquantes avec coefficients nuls.

PROPOSITION 7. *Toute matrice-colonne a à n éléments est une combinaison linéaire des matrices-colonnes $e_1, \dots, e_n$ introduites à l'aide de la formule* (4).

Cela se déduit de l'égalité suivante :

$$\begin{Vmatrix} a^1 \\ a^2 \\ \dots \\ a^n \end{Vmatrix} = a^1 \begin{Vmatrix} 1 \\ 0 \\ \dots \\ 0 \end{Vmatrix} + a^2 \begin{Vmatrix} 0 \\ 1 \\ 0 \\ \dots \\ 0 \end{Vmatrix} + \dots + a^n \begin{Vmatrix} 0 \\ \dots \\ 0 \\ 1 \end{Vmatrix}.$$

Les éléments de la matrice-colonne a sont les coefficients de la combinaison linéaire.

§ 2. Déterminants

Dans le chapitre I, on s'est déjà initié à la théorie des déterminants des matrices carrées d'ordre 2 et 3. L'objectif de ce paragraphe est de définir et d'étudier les déterminants des matrices carrées d'ordre quelconque. Il est d'usage dans ce cas de se servir du symbole suivant.

1. Le symbole $\sum$. En mathématiques, on est souvent obligé de considérer des sommes d'un grand nombre de termes de forme identique et ne différant que par les indices. Pour ces sommes on se sert de la notation suivante. Le symbole $\sum\limits_{k=1}^{n}$ suivi d'une expression munie de l'indice k désigne la somme de ces expressions pour toutes les valeurs de l'indice k de 1 à n, par exemple :

$$\sum_{k=1}^{n} a_k = a_1 + a_2 + \ldots + a_n, \quad \sum_{k=1}^{n} \alpha_k \beta_k = \alpha_1 \beta_1 + \ldots + \alpha_n \beta_n.$$

L'indice k s'appelle *indice de sommation*. Il va de soi que pour l'indice de sommation on peut utiliser toute autre lettre, par exemple :

$$\sum_{k=1}^{n} P_k = \sum_{s=1}^{n} P_s.$$

Citons quelques règles d'utilisation du symbole $\sum$ que le lecteur peut vérifier facilement.

PROPOSITION 1. 1) *Le facteur indépendant de l'indice de sommation peut être chassé de sous le symbole de sommation* :

$$\sum_{k=1}^{n} \alpha P_k = \alpha \sum_{k=1}^{n} P_k,$$

2)
$$\sum_{k=1}^{n} (P_k + Q_k) = \sum_{k=1}^{n} P_k + \sum_{k=1}^{n} Q_k.$$

La dernière formule est un cas particulier ($m = 2$) de l'assertion suivante.

PROPOSITION 2. *Deux symboles de sommation peuvent être permutés, c'est-à-dire que*

$$\sum_{k=1}^{n} \sum_{i=1}^{m} P_{ik} = \sum_{i=1}^{m} \sum_{k=1}^{n} P_{ik}.$$

En effet, les expressions P_{ik} intervenant sous les deux symboles de sommation dépendent de deux indices et peuvent être écrites sous forme de matrice à m lignes et n colonnes. Chacun des deux membres de l'égalité à démontrer est la somme de tous les éléments de la matrice : dans le premier membre, on additionne d'abord les éléments de chaque ligne et puis on calcule la somme des résultats obtenus, tandis que dans le second membre, on calcule d'abord la somme des éléments de chaque colonne et ensuite on additionne les sommes obtenues. La somme double $\sum_{i=1}^{n} \sum_{k=1}^{n} P_{ik}$ est notée $\sum_{i,k=1}^{n} P_{ik}$ si les deux indices parcourent les mêmes valeurs 1, ..., n.

L'application successive de la proposition 2 permet de démontrer que la somme multiple

$$\sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \cdots \sum_{i_p=1}^{n_p} P_{i_1 i_2 \ldots i_p}$$

est indépendante de l'ordre des symboles de sommation. Une telle somme est notée

$$\sum_{i_1 \ldots i_p} P_{i_1 i_2 \ldots i_p}$$

si tous les indices parcourent les mêmes valeurs.

Parfois, il nous faudra écrire la somme de tous les termes à l'exception d'un terme ou deux. S'il manque un seul terme d'indice j, on écrira : $\sum_{k \neq j} P_k$ et, s'il manque deux termes d'indices i et j, on écrira $\sum_{k \neq i, j} P_k$.

2. Définition du déterminant. Les déterminants ne sont définis que pour les matrices carrées. Le déterminant d'une matrice carrée est un nombre qui lui est associé et qui peut être calculé d'après ses éléments en se référant à la définition suivante.

DÉFINITION. 1) On appelle *déterminant* de la matrice d'ordre 1 l'élément de cette matrice.

2) On appelle *déterminant* de la matrice

$$A = \left\| \begin{matrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \multicolumn{4}{c}{\dotfill} \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{matrix} \right\|$$

d'ordre $n > 1$ le nombre

$$\det A = \sum_{k=1}^{n} (-1)^{k+1} a_{1k} M_k^1, \tag{1}$$

où M_k^1 est le déterminant de la matrice A_k^1 d'ordre $n - 1$ obtenue à partir de A en éliminant la première ligne et la k-ième colonne.

Pour désigner le déterminant de la matrice A, on utilise la notation det A ou, s'il est nécessaire d'écrire les éléments de la matrice, on les encadre de simples traits verticaux :

$$\left\| \begin{matrix} a_{11} & \cdots & a_{1n} \\ \multicolumn{3}{c}{\dotfill} \\ a_{n1} & \cdots & a_{nn} \end{matrix} \right\|.$$

A première vue, la définition du déterminant peut sembler peu efficace : en effet, on définit le déterminant de la matrice d'ordre n par l'intermédiaire des déterminants d'ordre $n - 1$, sans les avoir défini. Mais en réalité, il n'y a rien de gênant. Pour déterminer les nombres M_k^1 on peut se servir de la même formule puisqu'elle joue pour les matrices de tout ordre. On exprimera alors det A au moyen des déterminants d'ordre $n - 2$. En continuant le processus, on peut donc arriver à des matrices d'ordre 1 dont le déterminant est défini directement.

Appliquons notre définition aux matrices d'ordre 2 et 3. Pour la matrice

$$\left\| \begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{matrix} \right\|$$

on a $M_1^1 = a_{22}$, $M_2^1 = a_{21}$, et par suite,

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Pour la matrice

$$A = \left\| \begin{matrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{matrix} \right\|$$

il vient évidemment

$$M_1^1 = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}, \quad M_2^1 = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}, \quad M_3^1 = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

et

$$\det A = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{12}\begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix} + a_{13}\begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}.$$

Ces formules coïncident avec la définition des déterminants d'ordre 2 et 3, introduite au chapitre I.

Calculons à titre d'exemple le déterminant de la matrice

$$E_n = \left\| \begin{matrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{matrix} \right\|$$

appelée *matrice unité*. Pour cette matrice on a $M_1^1 = \det E_{n-1}$, où E_{n-1} est la matrice unité d'ordre $n-1$, et $a_{1k} = 0$ si $k \neq 1$. Par conséquent, $\det E_n = \det E_{n-1}$. En appliquant cette égalité $n-1$ fois, on obtient $\det E_n = \det E_1 = 1$.

Le nombre M_k^1 est appelé *mineur* de l'élément a_{1k}. Par analogie, on définit le mineur d'un élément quelconque a_{ij} comme le déterminant d'une matrice A_j^i obtenue à partir de la matrice initiale A en supprimant la ligne et la colonne à l'intersection desquelles se trouve l'élément a_{ij}, c'est-à-dire en supprimant la i-ième ligne et la j-ième colonne. Notons M_j^i le mineur de l'élément a_{ij}. On parle souvent des lignes et des colonnes du mineur ayant en vue les lignes et les colonnes de la matrice A_j^i. On se permettra cette liberté de langage car elle ne peut nous induire en erreur.

3. Propriétés des déterminants. Démontrons par récurrence les assertions suivantes.

PROPOSITION 3. *Pour toute matrice A d'ordre n on a la formule*

$$\det A = \sum_{i=1}^{n} (-1)^{i+1} a_{i1} M_1^i. \tag{2}$$

Cette formule est appelée *développement du déterminant suivant la première colonne*.

DÉMONSTRATION. Il est évident que pour les matrices d'ordre 2 la formule est vérifiée. Admettons que l'assertion est vraie pour les matrices d'ordre $n-1$ et démontrons qu'elle l'est pour toute matrice d'ordre n. Pour le faire, écrivons la formule (1) pour le déterminant de la matrice

d'ordre n en dégageant le premier terme de la somme :

$$\det A = a_{11}M_1^1 + \sum_{k=2}^{n} (-1)^{k+1}a_{1k}M_k^1.$$

Pour tout $k \geqslant 2$, la matrice A_k^1 comprend la première colonne (sans son premier élément) de la matrice A. En se servant de l'hypothèse de récurrence, on peut développer M_k^1 suivant cette colonne. Il faut seulement avoir en vue que la i-ième ligne de la matrice A devient la $(i-1)$-ième ligne de la matrice A_k^1 puisque A_k^1 ne contient pas la première ligne de la matrice A. Si donc $k \geqslant 2$, on a

$$M_k^1 = \sum_{i=2}^{n} (-1)^i a_{i1}M_{k1}^{1i},$$

où M_{k1}^{1i} est le déterminant d'une matrice d'ordre $n-2$ déduite de A_k^1 par suppression de la $(i-1)$-ième ligne et de la première colonne ou, ce qui revient au même, déduite de A par suppression de la première et de la i-ième ligne, et de la première et de la k-ième colonne.

En portant l'expression de M_k^1 obtenue on a

$$\det A = a_{11}M_1^1 + \sum_{k=2}^{n} \left\{(-1)^{k+1}a_{1k} \sum_{i=2}^{n} (-1)^i a_{i1}M_{k1}^{1i}\right\}.$$

Faisons entrer le facteur indépendant de i sous le deuxième symbole de sommation :

$$\det A = a_{11}M_1^1 + \sum_{k=2}^{n} \sum_{i=2}^{n} (-1)^{k+i+1}a_{1k}a_{i1}M_{k1}^{1i}.$$

Modifions l'ordre de sommation et chassons de facteur indépendant de k de sous le symbole de sommation :

$$\det A = a_{11}M_1^1 + \sum_{i=2}^{n} \left\{(-1)^{i+1}a_{i1} \sum_{k=2}^{n} (-1)^k a_{1k}M_{k1}^{1i}\right\}. \qquad (3)$$

Il n'est pas difficile de remarquer que la seconde somme est, par définition du déterminant, égale au mineur M_1^i. En effet,

$$M_1^i = \sum_{k=2}^{n} (-1)^k a_{1k}M_{1k}^{i1},$$

car la matrice A_1^i, dont le déterminant est M_1^i, ne renferme pas la première colonne de la matrice A et, de ce fait, tous les numéros des colonnes sont

décalés de 1. On peut maintenant écrire (3) sous la forme

$$\det A = a_{11} M_1^1 + \sum_{i=2}^{n} \{(-1)^{i+1} a_{i1} M_1^i\}.$$

L'expression obtenue coïncide avec l'égalité (2) qu'on s'est proposée de démontrer.

PROPOSITION 4. *Pour toute matrice carrée*, $\det A = \det {}^t A$.

Démontrons cette proposition par récurrence. Elle est évidente pour les matrices d'ordre 1. Admettons que la proposition est vraie pour les matrices d'ordre $n-1$ et démontrons-la pour les matrices d'ordre n. Soient A_j^1 la matrice déduite de A par suppression de la première ligne et de la j-ième colonne, et B_1^j la matrice obtenue à partir de ${}^tA$ par suppression de la j-ième ligne et de la première colonne. On voit immédiatement que $B_1^j = {}^tA_j^1$. Il ressort donc de l'hypothèse de récurrence que $\det B_1^j = \det A_j^1$, ou bien que le mineur associé à l'élément a_{1j} de la matrice A est égal au mineur associé à l'élément b_{j1} de la matrice ${}^tA$. Or $a_{1j} = b_{j1}$. Donc, le développement de $\det A$ suivant la première ligne coïncide avec le développement de $\det {}^tA$ suivant la première colonne.

De la proposition 4 résulte l'équivalence des lignes et des colonnes d'une matrice carrée. Plus précisément, toute assertion sur les déterminants des matrices qui est vraie pour les lignes est aussi vraie pour les colonnes, et inversement. En vertu de ce fait, il suffit de démontrer les propositions qui suivent pour les lignes seulement.

PROPOSITION 5. *Si l'on permute deux lignes (ou deux colonnes) quelconques dans une matrice carrée, son déterminant change de signe en gardant la valeur absolue.*

Démontrons cette assertion par récurrence en supposant que l'on permute deux lignes voisines de la matrice. Pour les matrices d'ordre 2 elle se vérifie directement. Admettons que l'assertion est vraie pour les matrices d'ordre $n-1$ et démontrons qu'elle est encore vraie pour les matrices d'ordre n.

Supposons que les numéros des lignes permutées sont k et $k+1$. Ecrivons le développement du déterminant suivant la première colonne en y dégageant deux termes correspondant aux lignes permutées :

$$\det A = (-1)^{k+1} a_{k1} M_1^k + (-1)^{k+2} a_{k+1,\,1} M_1^{k+1} + \sum_{i \neq k,\,k+1} (-1)^{i+1} a_{i1} M_1^i.$$

Procédons de façon analogue avec la matrice B déduite de A par permuta-

tion des lignes de numéros k et $k + 1$:

$$\det B = (-1)^{k+1} a_{k+1,1} N_1^k + (-1)^{k+2} a_{k1} N_1^{k+1} + \sum_{i \neq k, k+1} (-1)^{i+1} a_{i1} N_1^i.$$

Pour $i \neq k, k + 1$, les matrices de déterminants M_1^i et N_1^i contiennent les lignes de numéros k et $k + 1$, écrites dans l'ordre différent, les autres lignes étant les mêmes. Par hypothèse de récurrence on a donc $N_1^i = -M_1^i$, avec $i \neq k, k + 1$.

Les matrices dont les déterminants sont notés M_1^k et N_1^{k+1} coïncident : on les obtient en supprimant la $(k + 1)$-ième ligne de la matrice B ou, ce qui revient au même, la k-ième ligne de la matrice A. Donc, $M_1^k = N_1^{k+1}$. De façon analogue, $M_1^{k+1} = N_1^k$. En comparant maintenant $\det A$ et $\det B$, on voit qu'ils sont égaux en valeur absolue mais qu'ils ont des signes opposés.

Supposons maintenant que dans la matrice A d'ordre n on permute les lignes des numéros i et j, et soit pour fixer les idées $i < j$. Entre la i-ième et la j-ième lignes on a alors $j-i-1$ lignes. La permutation des i-ième et j-ième lignes ne s'effectue qu'avec les lignes voisines : on permute d'abord la j-ième ligne avec chacune des $j-i$ lignes situées au-dessus (dont la dernière est la i-ième) ; ensuite on fait passer la i-ième ligne à la place de la j-ième en la permutant avec chacune des $j-i-1$ lignes situées au-dessous d'elle. On réalisera ainsi un nombre impair de permutations de lignes voisines, soit $2(j - i) - 1$. Puisqu'avec chaque permutation le déterminant change de signe, il doit aussi modifier son signe avec la permutation des i-ième et j-ième lignes.

La propriété exprimée par la proposition 5 est appelée *antisymétrie* du déterminant par rapport aux lignes (resp. colonnes).

En utilisant la propriété d'antisymétrie par rapport aux lignes et aux colonnes, on est en mesure de démontrer que le déterminant peut être développé suivant toute ligne et toute colonne.

THÉORÈME 1. *Toute matrice A d'ordre n vérifie pour tout i, $1 \leqslant i \leqslant n$, la formule*

$$\det A = \sum_{k=1}^{n} (-1)^{k+i} a_{ik} M_k^i \tag{4}$$

et pour tout j, $1 \leqslant j \leqslant n$, la formule

$$\det A = \sum_{k=1}^{n} (-1)^{k+j} a_{kj} M_j^k. \tag{5}$$

Signalons que pour $i = 1$ la formule (4) devient la définition du déterminant, et que pour $j = 1$ la formule (5) coïncide avec le développement

(2) suivant la première colonne. Démontrons la formule (4) quand $i \geqslant 2$. A cet effet, transposons la i-ième ligne à la première place, de manière à ne pas déranger l'ordre des autres lignes. On le fera par des permutations successives de la i-ième ligne avec toutes les lignes situées au-dessus d'elle. Le nombre de ces lignes étant $i - 1$, on a $\det A = (-1)^{i-1} \det B$, où B est la matrice résultant de cette permutation. En développant $\det B$ suivant la première ligne (la i-ième ligne de la matrice A), on obtient

$$\det A = (-1)^{i-1} \sum_{k=1}^{n} (-1)^{k+1} a_{ik} N_k^1,$$

où N_k^1 est le déterminant de la matrice déduite de B par suppression de la première ligne et de la k-ième colonne ou, ce qui revient au même, déduite de A par suppression de la i-ième ligne et de la k-ième colonne. Donc, $N_k^1 = M_k^i$, ce qui démontre le développement (4).

La formule (5) peut être obtenue de la même façon par développement suivant la première colonne.

Proposition 6. *Si la i-ième colonne (resp. ligne) de la matrice A est une combinaison linéaire des matrices-colonnes (resp. matrices-lignes) p et q, autrement dit est de la forme $\alpha p + \beta q$, on a*

$$\det A = \alpha \det A_p + \beta \det A_q,$$

où les matrices A_p et A_q sont déduites de A par substitution respective de p et de q à la i-ième colonne (resp. ligne).

Pour le démontrer, il suffit de signaler qu'en vertu de la définition des opérations sur les matrices-colonnes on a pour tous les k $(1 \leqslant k \leqslant n)$ l'égalité $a_{ki} = \alpha p^k + \beta q^k$, où p^k et q^k désignent les éléments des matrices-colonnes p et q. En portant ces égalités dans le développement de $\det A$ suivant la i-ième colonne, on obtient

$$\det A = \sum_{k=1}^{n} (-1)^{k+i} a_{ki} M_i^k =$$
$$= \alpha \sum_{k=1}^{n} (-1)^{k+i} p^k M_i^k + \beta \sum_{k=1}^{n} (-1)^{k+i} q^k M_i^k,$$

ce qu'il fallait démontrer.

La propriété exprimée par cette proposition porte le nom de *linéarité du déterminant* par rapport à une colonne (resp. ligne). On peut évidemment représenter $\det A$ de façon analogue dans le cas où l'une des colonnes de la matrice A est la combinaison linéaire $\alpha_1 p_1 + \ldots + \alpha_s p_s$.

La linéarité du déterminant par rapport à ses colonnes est parfois formulée sous forme de deux propriétés :

1) *Si une colonne de la matrice est multipliée par un nombre, son déterminant est aussi multiplié par ce nombre.*

2) *Si une colonne de la matrice est la somme de deux colonnes, son déterminant est la somme des déterminants des matrices respectives.*

PROPOSITION 7. *Si dans la matrice A les colonnes (ou les lignes) sont linéairement dépendantes,* $\det A = 0$.

Signalons que si la matrice comporte une colonne (ou une ligne) nulle, son déterminant est nul. Cela résulte du développement suivant la colonne (resp. ligne). Admettons maintenant que la matrice ne comporte pas de colonne nulle.

Selon la proposition 2 du § 1, l'assertion à démontrer peut être formulée d'une autre manière : si une des colonnes (resp. lignes) de la matrice A est une combinaison linéaire des autres colonnes (resp. lignes), $\det A = 0$. Démontrons cette proposition. Partons d'un cas particulier.

Si A comporte deux colonnes identiques, leur permutation ne modifie pas la matrice, mais change le signe de son déterminant. Donc, $\det A = 0$.

Admettons maintenant que la j-ième colonne $\boldsymbol{a}_j$ est une combinaison linéaire des autres colonnes : $\boldsymbol{a}_j = \sum\limits_{k \neq j} \alpha_k \boldsymbol{a}_k$. (Certains coefficients α_k peuvent être nuls, c'est-à-dire que toutes les colonnes ne doivent pas forcément figurer dans cette combinaison linéaire.) En appliquant la propriété de linéarité par rapport aux colonnes, on a $\det A = \sum\limits_{k \neq j} \alpha_k \det A_k$, où A_k est la matrice déduite de A par substitution de la k-ième colonne à la j-ième. Dans cette matrice, la colonne $\boldsymbol{a}_k$ se répète deux fois ; donc $\det A_k = 0$ et la proposition est démontrée.

4. Transformations élémentaires. Calcul des déterminants. Le calcul des déterminants fait introduire et appliquer pour la première fois une notion importante de transformation élémentaire d'une matrice.

DÉFINITION. On appelle *transformations élémentaires d'une matrice* les transformations suivantes :

1) multiplication des éléments d'une ligne par un nombre non nul ;
2) addition d'une ligne à une autre ligne ;
3) permutation des lignes ;
4) mêmes transformations appliquées aux colonnes.

En combinant les transformations élémentaires de la première et de la deuxième forme, on est en mesure d'ajouter à toute ligne la combinaison linéaire des autres lignes.

PROPOSITION 8. *Le déterminant d'une matrice ne varie pas si à l'une quelconque de ses lignes (resp. colonnes) on ajoute la combinaison linéaire des autres lignes (resp. colonnes).*

Pour démontrer cette assertion, il faut appliquer la propriété de linéarité du déterminant par rapport aux lignes et tenir compte du fait que le déterminant dont les lignes sont linéairement dépendantes est nul.

Le déterminant de la matrice A peut être calculé ainsi. Si tous les éléments de sa première colonne sont nuls, $\det A = 0$. Mais si la première colonne contient des éléments non nuls, on choisit l'un d'eux (par exemple, le plus grand en valeur absolue). Posons que c'est a_{k1}. A chaque ligne, à l'exception de la k-ième, on ajoute la k-ième ligne multipliée par $-a_{i1}/a_{k1}$, où a_{i1} est le premier élément de la ligne subissant la transformation. Ainsi, dans la première colonne de la matrice transformée, tous les éléments sauf un s'annulent et il devient facile de développer son déterminant suivant la première colonne. On obtient

$$\det A = (-1)^{k+1} a_{k1} \det A',$$

où $\det A'$ est le mineur associé à l'élément a_{k1} dans la matrice transformée. Pour calculer $\det A'$ on utilise le même procédé. Après $n - 1$ cheminements (parfois moins) on obtient le déterminant.

Le procédé décrit a l'avantage par rapport aux autres car exige moins d'opérations arithmétiques. Pour trouver les déterminants de matrices pas trop grandes dont les éléments sont des nombres entiers ou des expressions algébriques simples, il est conseillé de l'utiliser en combinaison avec les autres procédés artificiels.

5. Mineurs d'ordre quelconque. Considérons une matrice A non nécessairement carrée et choisissons s lignes quelconques de numéros $i_1, \ldots, i_s$ et s colonnes quelconques de numéros $j_1, \ldots, j_s$. Soit $i_1 < i_2 < \ldots < i_s, j_1 < < j_2 < \ldots < j_s$.

DÉFINITION. On appelle *mineur d'ordre s* de la matrice A le déterminant d'une matrice carrée d'ordre s formée des éléments situés à l'intersection des lignes et colonnes choisies, autrement dit le nombre

$$L^{i_1 \ldots i_s}_{j_1 \ldots j_s} = \begin{Vmatrix} a^{i_1}_{j_1} & a^{i_1}_{j_2} & \ldots & a^{i_1}_{j_s} \\ a^{i_2}_{j_1} & a^{i_2}_{j_2} & \ldots & a^{i_2}_{j_s} \\ \ldots & \ldots & \ldots & \ldots \\ a^{i_s}_{j_1} & a^{i_s}_{j_2} & \ldots & a^{i_s}_{j_s} \end{Vmatrix}$$

Chaque matrice a autant de mineurs d'ordre donné s qu'il existe de possibilités de choisir les numéros $i_1, \ldots, i_s$ et $j_1, \ldots, j_s$. Il nous sera commode de dire que le mineur $L^{i_1 \ldots i_s}_{j_1 \ldots j_s}$ est situé à l'intersection des lignes de numéros $i_1, \ldots, i_s$ et des colonnes de numéros $j_1, \ldots, j_s$.

Si A est une matrice carrée, à chaque mineur $L^{i_1 \ldots i_s}_{j_1 \ldots j_s}$ d'ordre s on peut faire correspondre un mineur $M^{i_1 \ldots i_s}_{j_1 \ldots j_s}$ qui est égal au déterminant

d'une matrice carrée d'ordre $n - s$ obtenue de A en supprimant les lignes de numéros $i_1, \ldots, i_s$ et les colonnes de numéros $j_1, \ldots, j_s$, c'est-à-dire celles à l'intersection desquelles est situé le mineur $L_{j_1 \ldots j_s}^{i_1 \ldots i_s}$.

On appelle *cofacteur* d'un mineur son mineur associé multiplié par

$$(-1)^{i_1 + \ldots + i_s + j_1 + \ldots + j_s}.$$

On a la proposition suivante connue sous le nom de *théorème de Laplace*.

PROPOSITION 9. *Etant donné une matrice carrée, choisissons ses s lignes quelconques de numéros $i_1, \ldots, i_s$ et considérons tous les mineurs situés sur ses lignes. Le déterminant de la matrice considérée est alors égal à la somme des produits de chacun de ces mineurs par le cofacteur correspondant.*

Une assertion analogue peut être formulée pour les colonnes.

On ne fournira pas de démonstration à cette proposition vu qu'on ne s'y référera pas dans la suite. Signalons seulement que le développement du déterminant suivant une ligne est un cas particulier de la proposition 9 pour $s = 1$.

6. Expression du déterminant par les éléments de la matrice. On appelle *permutation* des nombres $1, \ldots, n$ ces nombres écrits dans un certain ordre. Par exemple, les nombres 1, 2 permettent de former deux permutations : (1, 2) et (2, 1). Une permutation quelconque des nombres $1, \ldots, n$ sera notée $(i_1, \ldots, i_n)$.

On dira que i_k est responsable du *dérangement de l'ordre* dans la permutation $(i_1, \ldots, i_n)$ s'il se trouve à gauche d'un plus petit nombre. Par exemple pour $n = 4$, chacun des nombres 2 et 3 est responsable d'un seul dérangement de l'ordre dans la permutation (2, 4, 3, 1) et le nombre 4 de deux. Ainsi donc, le nombre total de dérangements de l'ordre dans cette permutation est quatre. Le nombre de dérangements de l'ordre dans la permutation $(i_1, \ldots, i_n)$ sera noté $N(i_1, \ldots, i_n)$.

La permutation $(i_1, \ldots, i_n)$ est dite *paire* si $N(i_1, \ldots, i_n)$ est un nombre pair, et *impaire* dans le cas contraire.

Proposons-nous de démontrer la formule suivante :

$$\begin{vmatrix} a_{11} & \ldots & a_{1n} \\ \ldots & \ldots & \ldots \\ a_{n1} & \ldots & a_{nn} \end{vmatrix} = \sum_{(i_1 \ldots i_n)} (-1)^{N(i_1, \ldots, i_n)} a_{1i_1} a_{2i_2} \ldots a_{ni_n}. \qquad (6)$$

La somme dans le second membre de l'égalité est prise suivant toutes les permutations des nombres $1, \ldots, n$. Cela veut dire qu'à chaque permutation $(i_1, \ldots, i_n)$ des nombres $1, \ldots, n$ on fait correspondre un terme de la somme d'après la règle suivante : on prend un élément de la première ligne

et de la i_1-ième colonne, puis un élément de la deuxième ligne et de la i_2-ième colonne, etc. et on calcule leur produit qui contient donc un élément et un seul de chaque ligne et de chaque colonne. Leur produit est positif ou négatif suivant que la permutation correspondante est paire ou impaire.

DÉMONSTRATION. Raisonnons par récurrence pour démontrer la formule (6). Soit $n = 2$ et soit une matrice carrée

$$\begin{Vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{Vmatrix}.$$

Aux deux permutations (1, 2) et (2, 1) correspondent deux termes $(-1)^{N(1,\,2)}a_{11}a_{22}$ et $(-1)^{N(2,\,1)}a_{12}a_{21}$. Leur somme est égale à $a_{11}a_{22} - a_{12}a_{21}$, c'est-à-dire au déterminant même de la matrice donnée.

Admettons que la formule est vraie pour les matrices carrées d'ordre $n - 1$ et démontrons qu'elle l'est encore pour toute matrice carrée A d'ordre n. Le déterminant de A est défini par la formule (1) dont le k-ième terme contient le facteur M_k^1. Par hypothèse de récurrence

$$M_k^1 = \det A_k^1 = \sum_{(i_1 \ldots i_{n-1})} (-1)^{N(i_1,\, \ldots,\, i_{n-1})} a_{2i_1} \ldots a_{ni_{n-1}},$$

où tous les numéros $i_1, \ldots, i_{n-1}$ sont différents de k ; quant aux premiers indices affectant les facteurs, ce sont $2, \ldots, n$, car, en conservant les anciennes notations pour les éléments de la matrice A, on doit tenir compte du fait que la première ligne et la k-ième colonne ne figurent pas dans la matrice A_k^1.

On peut maintenant faire entrer le facteur $(-1)^{k+1}a_{1k}$ sous le symbole de sommation dans le k-ième terme de la formule (1) et l'écrire ainsi :

$$(-1)^{k+1}a_{1k}M_k^1 = \sum_{(i_1 \ldots i_{n-1})} (-1)^{N(i_1,\, \ldots,\, i_{n-1})+k+1} a_{1k}a_{2i_1} \ldots a_{ni_{n-1}}.$$

Les nombres $k, i_1, \ldots, i_{n-1}$ constituent une permutation des nombres $1, \ldots, n$, avec $N(k, i_1, \ldots, i_{n-1}) = N(i_1, \ldots, i_{n-1}) + k - 1$, car à droite de k il y a exactement $k - 1$ nombres inférieurs à k. Donc, $N(k, i_1, \ldots, i_{n-1})$ est de la même parité que $N(i_1, \ldots, i_{n-1}) + k + 1$. Il vient

$$(-1)^{k+1}a_{1k}M_k^1 = \sum_{(i_1 \ldots i_{n-1})} (-1)^{N(k,\, i_1,\, \ldots,\, i_{n-1})} a_{1k}a_{2i_1} \ldots a_{ni_{n-1}}.$$

Le second membre de cette expression réunit tous les termes de la somme (6) qui correspondent à des permutations ayant k à la première place. Etant donné que dans la somme (1) k parcourt toutes les valeurs de 1 à n, cette somme contient tous les termes de la somme (6) et eux seuls. La formule (6) est démontrée.

§ 3. Systèmes d'équations linéaires (cas spécial)

1. Position du problème. On appelle *système de m équations linéaires à n inconnues* $x^1, \ldots, x^n$ [*)] le système d'équations de la forme

$$\left.\begin{array}{l} a_1^1x^1 + a_2^1x^2 + \ldots + a_n^1x^n = b^1, \\ a_1^2x^1 + a_2^2x^2 + \ldots + a_n^2x^n = b^2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_1^mx^1 + a_2^mx^2 + \ldots + a_n^mx^n = b^m. \end{array}\right\} \quad (1)$$

Ecrivons les coefficients de ces équations sous forme de matrice

$$A = \left\| \begin{array}{cccc} a_1^1 & a_2^1 & \ldots & a_n^1 \\ \ldots & \ldots & \ldots & \ldots \\ a_1^m & a_2^m & \ldots & a_m^m \end{array} \right\|$$

appelée *matrice du système*. Les nombres figurant dans les seconds membres des équations constituent la matrice-colonne b appelée *matrice-colonne des termes constants*.

La matrice du système complétée par la matrice-colonne des termes constants est appelée *matrice complète du système* et est notée A^* dans ce chapitre :

$$A^* = \left\| \begin{array}{ccccc} a_1^1 & a_2^1 & \ldots & a_n^1 & b^1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_1^m & a_2^m & \ldots & a_n^m & b^m \end{array} \right\|.$$

Si les termes constants de toutes les équations sont nuls, le système est dit *homogène*.

DÉFINITION. Un ensemble de n nombres $\alpha^1, \ldots, \alpha^n$ est appelé *solution du système* (1) si toutes les équations du système deviennent des égalités numériques après substitution des nombres α^i, $i = 1, \ldots, n$, aux inconnues respectives x^i.

Proposons-nous de rechercher les solutions du système (1) sans faire de suppositions ni sur la nature des coefficients et des termes constants du système ni sur le nombre d'équations et d'inconnues. Aussi les situations qui peuvent se présenter, sont-elles les plus diverses. Le système peut même ne pas avoir de solution, comme par exemple le système

$$\begin{array}{l} x^1 + x^2 = 1, \\ x^1 + x^2 = 0 \end{array}$$

[*)] On écrit en haut les indices affectant les inconnues, pour pouvoir présenter dans la suite l'ensemble des inconnues sous forme de matrice-colonne.

(si la solution existait, $x^1 + x^2$ serait à la fois égal à 0 et à 1). Le système peut avoir une infinité de solutions, comme par exemple le système ($n = 2$, $m = 1$)

$$x^1 + x^2 = 0.$$

Les solutions de ce système sont tous les couples de nombres opposés. Quant aux exemples de systèmes présentant une solution unique, on les a rencontrés à maintes reprises dans le cours d'algèbre enseigné à l'école secondaire.

On dit que le système est *compatible* s'il a des solutions, et *incompatibles* s'il n'en a aucune.

2. Règle de Cramer. On examinera maintenant le cas le plus simple quand le nombre d'équations est égal à celui d'inconnues, soit $n = m$. On imposera de plus certaines conditions aux coefficients du système, car autrement on serait obligé d'examiner, par exemple, le système composé d'une même équation répétée n fois. Notre intention est d'étudier les systèmes dont toutes les équations sont, dans un certain sens, indépendantes. Déjà à l'école secondaire, on a largement utilisé le procédé suivant : on multipliait la première équation par un nombre α, la deuxième par β et on additionnait ces équations membre à membre. L'équation obtenue (naturellement appelée *combinaison linéaire* des équations initiales) est leur conséquence. Nous voulons donc qu'aucune équation du système ne soit une combinaison linéaire des autres. Si cette condition est satisfaite, on dira que les équations sont *linéairement indépendantes*.

Dans le cas où $m = n$, pour que les équations soient linéairement indépendantes il suffit d'exiger que le déterminant de la matrice du système soit différent de zéro. (Puisque $m = n$ et la matrice est carrée, on a toutes les raisons de parler de son déterminant.) En effet, remarquons qu'en multipliant une équation quelconque par un nombre, on multiplie la ligne correspondante de la matrice par ce nombre. En additionnant les équations, on additionne les lignes de la matrice. Il s'ensuit que si l'une des équations est une combinaison linéaire des autres, la ligne correspondante de la matrice du système est une combinaison linéaire des autres lignes. Il ressort alors de la proposition 7 du § 2 que dans ce cas le déterminant de la matrice est nul.

THÉORÈME 1 (RÈGLE DE CRAMER). *Le système de n équations à n inconnues*

$$\left.\begin{array}{c} a_1^1x^1 + \ldots + a_n^1x^n = b^1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_1^nx^1 + \ldots + a_n^nx^n = b^n \end{array}\right\} \quad (2)$$

possède une solution et une seule si le déterminant de la matrice du système

est non nul. Cette solution est définie par la formule

$$x^i = \frac{\Delta^i}{\Delta} \quad (\textit{pour tous les} \quad i = 1, \dots, n), \tag{3}$$

où Δ est le déterminant de la matrice du système et Δ^i le déterminant de la matrice déduite de la matrice du système par substitution de la colonne des termes constants à la i-ième colonne, c'est-à-dire

$$\Delta^i = \begin{vmatrix} a_1^1 & \dots & a_{i-1}^1 & b^1 & a_{i+1}^1 & \dots & a_n^1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_1^n & \dots & a_{i-1}^n & b^n & a_{i+1}^n & \dots & a_n^n \end{vmatrix}.$$

Pour démontrer le théorème considérons la matrice complète du système A^* et écrivons au-dessus d'elle l'une quelconque de ses lignes. Soit j le numéro de cette ligne. On obtient ainsi la matrice carrée $\bar{A}$ d'ordre $n + 1$. Cette matrice comporte deux lignes identiques et, par suite,

$$\det \bar{A} = \begin{vmatrix} a_1^j & \dots & a_n^j & b^j \\ a_1^1 & \dots & a_n^1 & b^1 \\ \dots & \dots & \dots & \dots \\ a_1^n & \dots & a_n^n & b^n \end{vmatrix} = 0.$$

On peut aussi calculer $\det \bar{A}$ en partant de la définition, soit :

$$\sum_{i=1}^{n} (-1)^{i+1} a_i^j M_i + (-1)^{n+1+1} b^j \det A = 0,$$

où M_i est le déterminant de la matrice déduite de la matrice complète A^* par suppression de la i-ième colonne. Donc, compte tenu de ce que $\Delta =$ $= \det A \neq 0$, on peut écrire

$$\frac{(-1)^{n+1}}{\Delta} \sum_{i=1}^{n} a_i^j (-1)^{i+1} M_i = b^j.$$

Si l'on fait entrer le facteur sous le symbole de sommation, cette égalité prend la forme

$$\sum_{i=1}^{n} a_i^j x^i = b^j,$$

où

$$x^i = \frac{(-1)^{n+i} M_i}{\Delta} \quad (i = 1, \dots, n).$$

L'ensemble des nombres $x^1, \dots, x^n$ ainsi défini vérifie, comme on le voit, la j-ième équation du système. Il est important que les nombres $x^1, \dots, x^n$ sont

indépendants de j et, de ce fait, vérifient toutes les équations du système, autrement dit, sont sa solution. On a ainsi démontré l'existence de la solution.

On donnera à x^i la forme voulue en mettant dans $\bar{A}$ la dernière colonne b à la i-ième place, c'est-à-dire en la permutant successivement avec les colonnes de numéros $n, n-1, \ldots, i+1$. Il faut faire en tout $n-i$ permutations. On a alors

$$x^i = \frac{(-1)^{n+i}(-1)^{n-i}\Delta^i}{\Delta} = \frac{\Delta^i}{\Delta}.$$

C'est justement la forme exigée pour x^i.

Il nous reste à démontrer l'unicité de la solution obtenue. Raisonnons par l'absurde. Soient deux solutions du système

$$\alpha^1, \ldots, \alpha^n \quad \text{et} \quad \beta^1, \ldots, \beta^n. \tag{4}$$

En utilisant les opérations avec les matrices-colonnes, on peut écrire (1) sous la forme (voir p. 125)

$$x^1 \begin{Vmatrix} a_1^1 \\ \cdots \\ a_1^n \end{Vmatrix} + \ldots + x^n \begin{Vmatrix} a_n^1 \\ \cdots \\ a_n^n \end{Vmatrix} = \begin{Vmatrix} b^1 \\ \cdots \\ b^n \end{Vmatrix} \tag{5}$$

ou de façon plus condensée $x^1 \boldsymbol{a}_1 + \ldots + x^n \boldsymbol{a}_n = \boldsymbol{b}$, où $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ sont les colonnes de la matrice du système, $\boldsymbol{b}$ la matrice-colonne des termes constants. Cette écriture du système est fort commode et on l'utilisera souvent.

La substitution des solutions (4) dans le système donne

$$\alpha^1 \boldsymbol{a}_1 + \ldots + \alpha^n \boldsymbol{a}_n = \boldsymbol{b},$$
$$\beta^1 \boldsymbol{a}_1 + \ldots + \beta^n \boldsymbol{a}_n = \boldsymbol{b}.$$

En retranchant membre à membre la seconde équation de la première, on obtient

$$(\alpha^1 - \beta^1)\boldsymbol{a}_1 + \ldots + (\alpha^n - \beta^n)\boldsymbol{a}_n = \boldsymbol{o}.$$

Si les solutions ne se confondent pas, au moins une des différences $\alpha^i - \beta^i$ est non nulle. Or cela signifie que les matrices-colonnes $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ sont linéairement dépendantes, ce qui contredit, en vertu de la proposition 7 du § 2, que $\det A \neq 0$. Le théorème est démontré.

Il vaut la peine de signaler que la coïncidence du nombre des équations avec celui des inconnues ne joue que vers la fin de la démonstration, ce qui nous permet de démontrer une assertion plus générale : si les colonnes de la matrice du système sont linéairement indépendantes, le système ne peut avoir deux solutions différentes.

3. Exemple. En appliquant la règle de Cramer, recherchons la condition à laquelle trois plans se coupent en un point. Soient trois plans définis dans

un repère cartésien par les équations

$$\left.\begin{aligned} A_1x + B_1y + C_1z + D_1 &= 0, \\ A_2x + B_2y + C_2z + D_2 &= 0, \\ A_3x + B_3y + C_3z + D_3 &= 0. \end{aligned}\right\} \qquad (6)$$

Les coordonnées du point d'intersection doivent vérifier simultanément les trois équations, c'est-à-dire être la solution du système (6). On voit qu'à la condition

$$\begin{vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \\ A_3 & B_3 & C_3 \end{vmatrix} \neq 0$$

il existe un point d'intersection et un seul. Rappelons que cette condition a été obtenue au § 3 du ch. II. La règle de Cramer n'entraîne que la suffisance de la condition, vu qu'elle représente, elle aussi, une condition suffisante d'existence et d'unicité de la solution.

Une autre interprétation géométrique de la règle de Cramer a été fournie pour le système de trois équations au point 6, § 3, ch. I. Il en résultait non seulement la condition d'unicité de la solution mais aussi les formules permettant de trouver cette solution.

L'étude des systèmes linéaires quelconques passe par celle des propriétés des matrices rectangulaires qui font l'objet du paragraphe suivant.

§ 4. Rang d'une matrice

1. Mineur principal. On a défini le mineur d'ordre r de la matrice A à la p. 137. Introduisons maintenant la définition suivante.

DÉFINITION. Etant donné une matrice à m lignes et n colonnes, on dit que son mineur d'ordre r est *principal* s'il est différent de zéro et si tous les mineurs d'ordre $r + 1$ sont nuls ou n'existent pas (dans le dernier cas, r coïncide avec le plus petit des nombres m ou n).

Il est évident qu'une matrice peut avoir plusieurs mineurs principaux. Tous les mineurs principaux sont du même ordre. En effet, si tous les mineurs d'ordre $r + 1$ sont nuls, sont également nuls tous les mineurs d'ordre $r + 2$ et, par suite, de tous les ordres supérieurs. L'assertion devient évidente si l'on applique la définition du déterminant à l'un quelconque des mineurs d'ordre $r + 2$; en effet, tous les mineurs associés aux éléments de sa première ligne sont des mineurs d'ordre $r + 1$ de la matrice considérée et, par suite, sont nuls.

On appelle *colonnes* et *lignes principales* les colonnes et les lignes à l'intersection desquelles est situé le mineur principal.

Définition. On appelle *rang* d'une matrice l'ordre du mineur principal, ou le plus grand ordre pour lequel il existe des mineurs non nuls. Si chaque élément de la matrice est nul, le rang de cette matrice est par définition égal à 0.

Le rang de la matrice A est noté $\mathrm{Rg}\,A$, ou rang A.

Trier un à un tous les mineurs pour rechercher le mineur principal est une opération exigeant d'énormes calculs si les dimensions de la matrice sont assez grandes. Il est beaucoup plus simple de calculer le rang de la matrice et son mineur principal en recourant aux transformations élémentaires.

Proposition 1. *Les transformations élémentaires ne modifient pas le rang d'une matrice.*

Démonstration. 1) Si on multiplie une ligne par le nombre $\lambda \neq 0$, le mineur principal devient multiplié par λ ou bien demeure inchangé. Aucun mineur nul ne devient différent de zéro.

2) Si tous les mineurs d'ordre $r + 1$ sont nuls, l'addition des lignes ne rend aucun d'eux différent de zéro. En effet, le mineur obtenu dans cette transformation est égal soit à la somme algébrique de deux mineurs d'ordre $r + 1$ de la matrice initiale (si à une ligne du mineur on ajoute une autre ligne qui n'y figure pas), soit à la somme d'un mineur d'ordre $r + 1$ et d'un déterminant de la matrice ayant deux lignes identiques (si à une ligne du mineur on ajoute une ligne qui y figure). Ces considérations entraînent que le rang de la matrice ne peut augmenter. Il va de soi qu'il ne peut non plus diminuer, car autrement la soustraction des lignes, qui est une transformation inverse, l'aurait fait augmenter.

3) Si on permute les lignes de la matrice, le mineur peut modifier son signe (si les deux lignes permutées lui appartiennent) ou devenir égal, au signe près, à un autre mineur de la même matrice (s'il contient une seule ligne des lignes permutées), ou bien ne pas se modifier du tout. Il est évident que dans ce cas l'ordre du mineur principal reste inchangé.

4) On démontre de façon analogue que les transformations élémentaires des colonnes ne changent pas le rang de la matrice.

Il sera préférable pour nous de considérer les transformations élémentaires des lignes de la matrice, vu leur étroite liaison avec les transformations des systèmes d'équations linéaires. Pour le système de m équations à n inconnues

$$\left.\begin{array}{c} a_{11}x^1 + \ldots + a_{1n}x^n = b_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x^1 + \ldots + a_{mn}x^n = b_m \end{array}\right\} \qquad (1)$$

la matrice complète prend la forme

$$A^* = \left\| \begin{matrix} a_{11} & \dots & a_{1n} & b_1 \\ \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} & b_m \end{matrix} \right\| .$$

La permutation des lignes de cette matrice équivaut au changement de l'ordre des équations du système. La multiplication d'une ligne par le nombre $\lambda \neq 0$ équivaut à la multiplication de l'équation correspondante par ce nombre. Enfin, ajouter une ligne de la matrice A^* à une autre revient à additionner les équations correspondantes du système. Dans toutes les transformations l'ensemble des solutions du système ne varie évidemment pas. On a démontré ainsi la

PROPOSITION 2. *A chaque transformation élémentaire des lignes de la matrice complète correspond une transformation du système d'équations linéaires en un système équivalent.*

Montrons maintenant comment obtenir le mineur principal d'une matrice à l'aide des transformations élémentaires.

2. Obtention de la forme simplifiée d'une matrice. Soit une matrice A de type (m, n). Si tous ses éléments sont nuls, le rang de la matrice est 0 et le mineur principal n'existe pas. Si ce n'est pas le cas, soit j_1 le numéro de la première colonne contenant des éléments non nuls et soit $a_{i_1j_1}$ un élément non nul de cette colonne. Portons la i_1-ième ligne à la première place et divisons-la par $a_{i_1j_1}$. Notons a'_{ij} les éléments de la matrice transformée. Il vient alors $a'_{1j_1} = 1$. Effectuons les transformations élémentaires des lignes pour que tous les autres éléments de la j_1-ième colonne s'annulent. Il faut pour cela soustraire la première ligne multipliée par a'_{kj_1} de chaque ligne de numéro (nouveau) $k \neq 1$.

Après ces transformations la matrice prend la forme

$$A_1 = \left\| \begin{matrix} 0 & \dots & 0 & 1 \\ 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 0 \end{matrix} \right\| U_1 ,$$

où U_1 est une matrice à m lignes et $n - j_1$ colonnes.

Si les $m - 1$ dernières lignes de la matrice A_1 sont nulles, on achève les transformations. Sinon, soit j_2 le numéro de la première colonne à élément non nul dans l'une des $m - 1$ dernières lignes. Transportons la ligne avec cet élément non nul à la deuxième place et divisons-la par cet élément. Notons a''_{ij} les éléments de la matrice transformée. Alors $a''_{2j_2} = 1$.

Effectuons les transformations élémentaires nécessaires pour que tous

les autres éléments de la j_2-ième colonne deviennent nuls. Pour le faire, il faut soustraire la deuxième ligne multipliée par a''_{kj_2} de chaque ligne de numéro $k \neq 2$. Ceci étant, les j_1 premières colonnes de la matrice A_1 ne varient pas, car les éléments situés à l'intersection de ces colonnes avec la deuxième ligne sont des zéros. La matrice prend maintenant la forme

$$A_2 = \left\| \begin{array}{cccccccc|c} 0 & \dots & 0 & 1 & \underline{*} & \underline{\dots} & \underline{*} & 0 & \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 1 & U_2 \\ \multicolumn{8}{c|}{\dots\dots\dots\dots\dots\dots\dots\dots} & \\ 0 & \dots & 0 & 0 & 0 & \dots & 0 & 0 & \end{array} \right\|,$$

où U_2 est une matrice à m lignes et $n - j_1 - j_2$ colonnes ; les étoiles désignent des éléments dont on ne peut rien dire.

Si dans les $m - 2$ dernières lignes il y a des éléments non nuls, on effectue des transformations analogues qui permettent de passer de la j_3-ième colonne à une colonne dont tous les éléments sont des zéros à l'exception du troisième élément égal à l'unité. A gauche de cette colonne, dans les $m - 2$ dernières lignes on ne trouve que des éléments nuls, de sorte que les colonnes déjà transformées ne changent pas.

On poursuit ces transformations jusqu'à ce que les $m - r$ dernières lignes de la matrice A_r ne présentent que des zéros ou que toutes les lignes soient épuisées.

On a ainsi la

PROPOSITION 3. *Toute matrice de type* (m, n) *peut être transformée en une matrice dont r colonnes coïncident avec les r premières colonnes de la matrice unité d'ordre m. Les* $m - r$ *dernières lignes sont constituées de zéros si* $r < m$.

On appellera *forme simplifiée* la forme de la matrice décrite dans cette proposition. Les matrices de cette forme seront dites tout simplement *simplifiées*.

Considérons une matrice de forme simplifiée. Son mineur situé à l'intersection des r premières lignes et des colonnes $j_1, \dots, j_r$ est égal à l'unité. Il n'y a évidemment pas de mineurs non nuls d'ordre supérieur. Donc, c'est un mineur principal, et le rang de la matrice est r.

Si l'on peut passer de la matrice A à une matrice simplifiée de rang r, cela signifie que la matrice A est également de rang r. En effet, les transformations élémentaires ne modifient pas le rang de la matrice. Pour mineur principal de la matrice A on peut prendre le mineur situé à l'intersection des colonnes de numéros $j_1, \dots, j_r$ et des lignes qui, toutes les permutations étant effectuées, prennent les numéros $1, \dots, r$ dans la matrice simplifiée. Cela découle du fait qu'en transformant la matrice A on n'a ajouté aux

lignes de ce mineur aucune ligne qui n'y figure pas. Puisque le mineur obtenu n'est pas nul, il en est de même du mineur initial.

Remarquons que la matrice dont il s'agit dans la proposition 3 résulte des transformations élémentaires effectuées sur les lignes. Cette remarque est importante à cause de la proposition 2. La méthode utilisée dans la démonstration est appelée *méthode de Gauss* (ou de *Gauss-Jordan*).

Considérons une matrice carrée dont le déterminant est différent de zéro. Toutes ses colonnes sont principales et, par suite, sa forme simplifiée représente une matrice unité. D'où le

COROLLAIRE. *Toute matrice carrée de déterminant non nul peut être réduite par transformations élémentaires à une matrice unité.*

La recherche du mineur principal dans la démonstration de la proposition 3 aboutit à un ensemble des premières colonnes renfermant le mineur principal. Pour une matrice de forme simplifiée ce fait est sans importance.

PROPOSITION 4. *Soit A une matrice de type (m, n). Quel que soit le mineur principal de cette matrice, il existe des transformations élémentaires des lignes de A qui permettent de passer des colonnes du mineur à celles de la matrice unité. Si* $\mathrm{Rg}\, A = r < m$, *les $m - r$ dernières lignes sont nulles.*

DÉMONSTRATION. Supposons que le mineur principal est situé à l'intersection des lignes $i_1, \ldots, i_r$ et des colonnes $j_1, \ldots, j_r$. Portons ces colonnes à la place des r premières et procédons comme dans la démonstration de la proposition 3, à la seule différence qu'on doit choisir un élément non nul de la colonne suivante parmi les éléments des lignes de numéros $i_1, \ldots, i_r$. Cet élément existe toujours, sinon le mineur principal s'annulerait par transformations élémentaires de ses lignes. Les transformations étant effectuées, il faut ramener les colonnes à leurs places. On peut d'ailleurs ne pas toucher aux colonnes et modifier quelque peu la méthode de réduction.

3. Théorème du mineur principal. L'exposé qui suit s'appuie dans une large mesure sur une assertion connue sous le nom de *théorème du mineur principal*.

THÉORÈME 1. *Chaque colonne (resp. ligne) de la matrice est une combinaison linéaire des colonnes (resp. lignes) principales.*

Pour démontrer l'assertion du théorème relative aux lignes, revenons aux transformations élémentaires utilisées dans la démonstration de la proposition 4 pour réduire la matrice à sa forme simplifiée. Outre les permutations des lignes et la multiplication des lignes principales par des nombres, on ajoutait à une ligne quelconque de la matrice l'une de ses lignes principales multipliée par un nombre. Les lignes principales se remplaçaient dans ce cas par des combinaisons linéaires des lignes principales, et aux lignes non principales venaient s'ajouter les combinaisons des lignes principales.

Dans une matrice simplifiée, les lignes non principales sont nulles. Cela signifie que pour chaque ligne non principale de la matrice initiale on a trouvé une combinaison linéaire de ses lignes principales telle que leur somme est la ligne nulle. L'assertion est démontrée.

L'assertion du théorème relative aux colonnes sera aussi démontrée si l'on applique l'assertion déjà démontrée pour les lignes à la matrice transposée en notant préalablement que le mineur principal de la matrice est encore un mineur principal de sa transposée.

On a affirmé dans la proposition 7 du § 2 que le déterminant de la matrice dont les colonnes sont linéairement dépendantes est nul. Le théorème du mineur principal entraîne une assertion réciproque.

PROPOSITION 5. *Si A est une matrice carrée et* $\det A = 0$, *il existe au moins une colonne* (*resp. ligne*) *qui est une combinaison linéaire des autres colonnes* (*resp. lignes*).

En effet, la condition $\det A = 0$ signifie que $\operatorname{Rg} A \leqslant n - 1$ où n est l'ordre de la matrice. Donc, il existe au moins une colonne (resp. ligne) qui ne coupe pas le mineur principal. Or cette colonne (resp. ligne) s'exprime linéairement par les colonnes (resp. lignes) sur lesquelles est situé le mineur principal.

THÉORÈME 2. (THÉORÈME DU RANG D'UNE MATRICE). *Le rang d'une matrice A est égal au nombre maximal de ses colonnes linéairement indépendantes.*

Si $\operatorname{Rg} A = 0$, toutes les colonnes sont nulles et aucune colonne n'est linéairement indépendante. Soit $\operatorname{Rg} A = r > 0$. Montrons que dans A il existe r colonnes linéairement indépendantes. En effet, considérons une matrice carrée A' d'ordre r extraite de la matrice A, dont le déterminant est le mineur principal. Chaque colonne de A' représente une partie de la colonne correspondante de A.

Si les colonnes de A renfermant le mineur principal étaient linéairement dépendantes, il en serait de même des colonnes de A' et le mineur principal serait nul.

Démontrons maintenant que toute famille de p colonnes de la matrice A est liée si $p > r$. Soit B une matrice composée de p colonnes quelconques de A. On a $\operatorname{Rg} B \leqslant r$. En effet, chaque mineur de la matrice B est le mineur de la matrice A et par suite, il n'y a dans B aucun mineur non nul dont l'ordre est strictement supérieur à r. Ainsi, $\operatorname{Rg} B < p$ et au moins une des colonnes de la matrice B n'appartient pas à son mineur principal.

Cette colonne s'exprime linéairement au moyen des autres colonnes. Le théorème est ainsi démontré.

On démontre de la même façon que le rang de la matrice est égal au nombre maximal de lignes linéairement indépendantes.

COROLLAIRE. *Le nombre maximal de lignes linéairement indépendantes dans une matrice est égal au nombre maximal de colonnes linéairement indépendantes de cette matrice.*

§ 5. Théorie générale des systèmes linéaires

1. Condition de compatibilité. Le théorème du mineur principal permet de formuler une condition simple et efficace de compatibilité du système d'équations linéaires

$$\left.\begin{array}{l} a_1^1 x^1 + \ldots + a_n^1 x^n = b^1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_1^m x^1 + \ldots + a_n^m x^n = b^m, \end{array}\right\} \quad (1)$$

qui porte le nom de *théorème de Kronecker-Capelli.*

THÉORÈME 1. *Le système* (1) *possède au moins une solution si et seulement si le rang de la matrice du système A est égal au rang de la matrice complète A^*.*

Pour le démontrer, mettons le système (1) sous une autre forme en utilisant les définitions des opérations sur les matrices-colonnes. Il vient

$$x^1 \left\| \begin{matrix} a_1^1 \\ \ldots \\ a_1^m \end{matrix} \right\| + \ldots + x^n \left\| \begin{matrix} a_n^1 \\ \ldots \\ a_n^m \end{matrix} \right\| = \left\| \begin{matrix} b^1 \\ \ldots \\ b^m \end{matrix} \right\| . \quad (2)$$

1) Si la solution existe, l'égalité (2) signifie que la matrice-colonne des termes constants est une combinaison linéaire des colonnes de la matrice du système. Donc, l'adjonction de cette matrice-colonne n'augmente pas le nombre total de colonnes linéairement indépendantes, et $\mathrm{Rg}\, A^* = \mathrm{Rg}\, A$.

2) Soit $\mathrm{Rg}\, A^* = \mathrm{Rg}\, A$. Dans ce cas, le mineur principal de la matrice A l'est encore de la matrice A^*. Cela signifie que la matrice-colonne des termes constants est une combinaison linéaire des colonnes de la matrice A qui renferment le mineur principal. Selon la proposition 6 du § 1, la matrice-colonne des termes constants est dans ce cas une combinaison linéaire de toutes les colonnes de la matrice A. Les coefficients de cette combinaison linéaire représentent la solution du système (1).

COROLLAIRE.. *Le système* (1) *est incompatible si et seulement si la matrice simplifiée associée à la matrice complète A^* contient la ligne* $\| 0 \ldots 0 \ 1 \|$. (On suppose que les colonnes n'ont pas été permutées.)

DÉMONSTRATION. Si $\mathrm{Rg}\, A^* > \mathrm{Rg}\, A$, tout mineur principal de la matrice A^* doit contenir la dernière colonne. Or si la dernière colonne est principale, la matrice simplifiée associée à la matrice A^* doit comporter la ligne

$\|0 \dots 0\ 1\|$. Inversement, si la matrice simplifiée associée à la matrice A^* contient une telle ligne, sa dernière colonne ne peut être une combinaison linéaire des autres colonnes, et le système correspondant à la matrice simplifiée est donc incompatible. Or ce système est équivalent au système initial.

A titre d'exemple, déduisons du théorème de Kronecker-Capelli la condition de parallélisme de deux droites non confondues d'un même plan. Les droites

$$\left.\begin{aligned} A_1x + B_1y + C_1 &= 0, \\ A_2x + B_2y + C_2 &= 0 \end{aligned}\right\} \tag{3}$$

sont parallèles si elles n'ont aucun point commun, c'est-à-dire si

$$\operatorname{Rg}\begin{Vmatrix} A_1 & B_1 \\ A_2 & B_2 \end{Vmatrix} < \operatorname{Rg}\begin{Vmatrix} A_1 & B_1 & C_1 \\ A_2 & B_2 & C_2 \end{Vmatrix}.$$

On peut exprimer l'égalité des rangs de la matrice du système et de la matrice complète en interprétant le rang comme le nombre maximal de lignes linéairement indépendantes. On aboutit ainsi à la proposition 1 formulée plus loin et connue sous le nom de *théorème de Fredholm*. Quoique s'adaptant difficilement à l'étude des systèmes d'équations, ce théorème est d'une grande importance théorique.

Transposons la matrice A du système (1) et considérons le système de n équations linéaires

$$\left.\begin{aligned} a_1^1 y_1 + \dots + a_1^m y_m &= 0, \\ &\dots\dots\dots \\ a_n^1 y_1 + \dots + a_n^m y_m &= 0 \end{aligned}\right\} \tag{4}$$

à m inconnues, dont la matrice est ${}^tA$ et dont les seconds membres sont nuls. On dit que ce système est *homogène adjoint* du système (1).

PROPOSITION 1 (THÉORÈME DE FREDHOLM). *Pour que le système d'équations linéaires* (1) *soit compatible il faut et il suffit que toute solution du système homogène adjoint* (4) *vérifie l'équation*

$$b^1 y_1 + \dots + b^m y_m = 0, \tag{5}$$

où $b^1, \dots, b^m$ *sont les termes constants des équations* (1).

Démontrons que la condition est suffisante. Si la relation (5) est vérifiée par toutes les solutions du système (4), le système d'équations

$$\left.\begin{aligned} a_1^1 y_1 + \dots + a_1^m y_m &= 0, \\ &\dots\dots\dots \\ a_n^1 y_1 + \dots + a_n^m y_m &= 0, \\ b^1 y_1 + \dots + b^m y_m &= 1 \end{aligned}\right\} \tag{6}$$

ne peut avoir de solutions. Il en ressort que le rang de la matrice de ce système est strictement inférieur au rang de la matrice complète. Transposons ces deux matrices et tenons compte de ce que cette opération ne change pas leur rang. On obtient

$$\mathrm{Rg} \begin{Vmatrix} a_1^1 & \dots & a_n^1 & b^1 \\ \dots & \dots & \dots & \dots \\ a_1^m & \dots & a_n^m & b^m \\ 0 & \dots & 0 & 1 \end{Vmatrix} > \mathrm{Rg} \begin{Vmatrix} a_1^1 & \dots & a_n^1 & b^1 \\ \dots & \dots & \dots & \dots \\ a_1^m & & a_n^m & b^m \end{Vmatrix}.$$

Le théorème du rang d'une matrice implique maintenant que la matrice-ligne $\|0 \dots 0 \ 1\|$ n'est pas une combinaison linéaire des lignes de la matrice

$$A^* = \begin{Vmatrix} a_1^1 & \dots & a_n^1 & b^1 \\ \dots & \dots & \dots & \dots \\ a_1^m & \dots & a_n^m & b^m \end{Vmatrix}$$

et, par suite, cette ligne ne peut pas figurer dans la matrice simplifiée associée à cette matrice. Il s'ensuit, compte tenu du corollaire au théorème de Kronecker-Capelli, que le système (1) est compatible.

Démontrons maintenant que la condition est nécessaire. Raisonnons par l'absurde en supposant que cette condition n'est pas satisfaite. Il existe alors une solution $y_1, \dots, y_m$ du système (4) pour laquelle $b_1y_1 + \dots + b_my_m = p \neq 0$. Multiplions les équations du système (1) par les nombres respectifs $y_1, \dots, y_m$ et additionnons-les. On obtient l'équation

$$0x^1 + \dots + 0x^n = p,$$

qui n'a aucune solution. Si le système (1) avait des solutions, elles devraient vérifier cette équation. Donc, le système (1) est incompatible : contradiction.

A titre d'exemple, appliquons le théorème de Fredholm pour déduire la condition de parallélisme de deux droites non confondues d'un même plan. Le système (3) n'a pas de solutions s'il existe des nombres y_1 et y_2 pour lesquels $y_1A_1 + y_2A_2 = 0$, $y_1B_1 + y_2B_2 = 0$ mais $y_1C_1 + y_2C_2 \neq 0$. On constate facilement que y_1 et y_2 ne sont pas nuls. On peut donc poser $-y_2/y_1 = \lambda$ et écrire la condition obtenue sous la forme : il existe un nombre λ tel que $A_1 = \lambda A_2$, $B_1 = \lambda B_2$ et $C_1 \neq \lambda C_2$.

2. Recherche des solutions. Soit donné un système compatible de m équations linéaires à n inconnues. Notons r le rang de la matrice du système. Vu que le rang de la matrice complète est aussi égal à r, on peut choisir le mineur principal de la matrice complète de manière qu'il soit contenu dans la matrice du système. Par des transformations élémentaires des lignes, réduisons la matrice complète à la matrice simplifiée (proposition 3 du § 4). Selon la proposition 2 du § 4, le système donné d'équations linéai-

res se transforme alors en un système équivalent de r équations linéairement indépendantes.

Admettons pour plus de commodité que le mineur principal est situé sur les r premières colonnes. Le système transformé peut alors être écrit sous la forme suivante :

$$\left.\begin{aligned} x^1 &= \bar{b}^1 - (\bar{a}^1_{r+1}x^{r+1} + \ldots + \bar{a}^1_n x^n), \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x^r &= \bar{b}^r - (\bar{a}^r_{r+1}x^{r+1} + \ldots + \bar{a}^r_n x^n), \end{aligned}\right\} \tag{7}$$

où $\bar{a}^i_j$ et $\bar{b}^i$ sont les éléments de la matrice complète transformée. Dans les premiers membres des égalités, on a écrit les inconnues correspondant aux colonnes du mineur principal choisi, appelées *inconnues principales*. Les autres inconnues, dites *paramétriques* ou non principales, sont portées dans les seconds membres des égalités.

Quelles que soient les inconnues paramétriques, les formules (7) permettent de trouver les inconnues principales qui constituent toutes ensemble la solution du système (1). On voit aisément que ce procédé nous permet d'obtenir toutes les solutions du système.

On aurait pu s'arrêter aux formules (7), toutefois on donnera plus loin une description de l'ensemble des solutions du système d'équations linéaires qui est beaucoup plus simple et parlante, ainsi que présentant une importance de principe.

3. Système homogène associé. Associons au système d'équations (1) un *système homogène* ayant la même matrice des coefficients, autrement dit le système

$$\left.\begin{aligned} a^1_1x^1 + \ldots + a^1_n x^n &= 0, \\ &\ldots\ldots\ldots\ldots\ldots \\ a^m_1x^1 + \ldots + a^m_n x^n &= 0. \end{aligned}\right\} \tag{8}$$

On l'appelle *système homogène associé à* (1).

PROPOSITION 2. *Si $y^1_0, \ldots, y^n_0$ est une solution du système* (1), *$y^1, \ldots, y^n$ est aussi une solution du système* (1) *si et seulement s'il existe une solution $x^1, \ldots, x^n$ du système homogène associé à* (1) *telle que $y^i = y^i_0 + x^i$ pour tout $i = 1, \ldots, n$.*

DÉMONSTRATION. 1) Soient $a_1, \ldots, a_n$ les colonnes de la matrice du système (1), et b la matrice-colonne de ses seconds membres. Soient de plus $y^1_0, \ldots, y^n_0$ la solution :

$$y^1_0 a_1 + \ldots + y^n_0 a_n = b, \tag{9}$$

et $x^1, \ldots, x^n$ la solution du système homogène associé à (1) :

$$x^1 a_1 + \ldots + x^n a_n = o. \tag{10}$$

En additionnant membre à membre les égalités (9) et (10), on obtient

$$(y_0^1 + x^1)a_1 + \dots + (y_0^n + x^n)a_n = b,$$

c'est-à-dire que les nombres $y_0^1 + x^1, \dots, y_0^n + x^n$ vérifient le système (1).

2) Soit $y^1, \dots, y^n$ une solution du système (1) :

$$y^1a_1 + \dots + y^na_n = b. \tag{11}$$

En retranchant (9) de (11), on a

$$(y^1 - y_0^1)a_1 + \dots + (y^n - y_0^n)a_n = o.$$

Cela signifie que les nombres $x^1 = y^1 - y_0^1, \dots, x^n = y^n - y_0^n$ vérifient le système homogène associé. La proposition est démontrée.

En écrivant la solution sous forme de matrice-colonne à n éléments on peut formuler la proposition 2 ainsi : si y_0 est une solution du système (1), il en est de même de y si et seulement s'il existe une solution x du système homogène associé telle que $y = y_0 + x$.

Dans la suite, pour abréger on utilisera cette écriture.

4. Ensemble des solutions d'un système homogène. Tout système homogène d'équations linéaires est compatible. Il a pour solution 0, ..., 0 que l'on appelle *solution triviale*. Tous les résultats précédents obtenus pour les systèmes d'équations linéaires sont évidemment vrais pour les systèmes homogènes. En particulier, sont vraies les formules (7) qui pour le système homogène prennent la forme

$$\left.\begin{array}{l} x^1 = -(\bar{a}^1_{r+1}x^{r+1} + \dots + \bar{a}^1_nx^n), \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ x^r = -(\bar{a}^r_{r+1}x^{r+1} + \dots + \bar{a}^r_nx^n). \end{array}\right\} \tag{7'}$$

En effet, dans les formules (7) on a $\bar{b}^1 = 0, \dots, \bar{b}^r = 0$, vu que les opérations élémentaires sur les lignes transforment la matrice-colonne nulle des termes constants en une matrice-colonne nulle.

L'ensemble des solutions d'un système homogène possède deux propriétés importantes qu'exprime la proposition suivante.

PROPOSITION 3. *Si les matrices-colonnes x_1 et x_2 sont des solutions du système homogène, leur somme $x_1 + x_2$ est aussi une solution de ce système. Le produit de la solution d'un système homogène par un nombre est encore une solution de ce système.*

DÉMONSTRATION. Etant donné que

$$x_1^1a_1 + \dots + x_1^na_n = o \quad \text{et} \quad x_1^1a_1 + \dots + x_2^na_n = o,$$

il vient

$$(x_1^1 + x_2^1)a_1 + \dots + (x_1^n + x_2^n)a_n = o$$

et

$$\lambda x_1^1 a_1 + \ldots + \lambda x_1^n a_n = o,$$

quel que soit λ. La proposition est démontrée.

Dans la démonstration des deux propositions suivantes, on utilise les formules (7′). Aussi la numération des variables adoptée y reflète-t-elle l'hypothèse faite pour (7) et (7′) que le mineur principal de la matrice du système est situé sur les r premières colonnes.

PROPOSITION 4. *Si le rang de la matrice d'un système homogène est r, le système possède $n - r$ solutions linéairement indépendantes.*

Pour le démontrer, associons aux inconnues paramétriques $n - r$ ensembles de valeurs :

1) $x^{r+1} = 1, \quad r^{r+2} = 0, \quad \ldots, \quad \ldots, \quad x^n = 0;$

2) $x^{r+1} = 0, \quad x^{r+2} = 1, \quad x^{r+3} = 0, \quad \ldots, \quad x^n = 0;$

....

$n - r$) $x^{r+1} = 0, \quad \ldots, \quad \ldots, \quad x^{n-1} = 0, \quad x^n = 1.$

Pour chaque ensemble de valeurs des inconnues paramétriques cherchons les valeurs correspondantes des inconnues principales. Ecrivons les solutions obtenues sous forme de matrices-colonnes :

$$x_1 = \left\| \begin{array}{c} x_1^1 \\ \ldots \\ x_1^r \\ 1 \\ 0 \\ \ldots \\ 0 \end{array} \right\|, \quad x_2 = \left\| \begin{array}{c} x_2^1 \\ \ldots \\ x_2^r \\ 0 \\ 1 \\ \ldots \\ 0 \end{array} \right\|, \quad \ldots, \quad x_{n-r} = \left\| \begin{array}{c} x_{n-r}^1 \\ \ldots \\ x_{n-r}^r \\ 0 \\ \ldots \\ 0 \\ 1 \end{array} \right\|. \qquad (12)$$

Ces solutions sont linéairement indépendantes. En effet, soit la matrice dont les colonnes sont $x_1, \ldots, x_{n-r}$. Elle possède un mineur d'ordre $n - r$ (situé sur les $n - r$ dernières lignes) égal à l'unité. Par conséquent, son rang est $n - r$ et toutes ses colonnes sont linéairement indépendantes.

L'ensemble des solutions (12) est appelé *système fondamental normal de solutions*. D'une façon générale on appelle *système fondamental de solutions* tout ensemble de $n - r$ solutions linéairement indépendantes.

PROPOSITION 5. *Soit $x_1, \ldots, x_{n-r}$ un ensemble fondamental de solutions d'un système d'équations linéaires homogènes. Toute solution x du système est alors une combinaison linéaire des solutions $x_1, \ldots, x_{n-r}$.*

Pour le démontrer, considérons la matrice X dont les colonnes sont les solutions x et $x_1, \ldots, x_{n-r}$. Le rang de X est au moins égal à $n - r$, car cette

matrice comprend $n - r$ colonnes linéairement indépendantes. D'autre part, il est au plus égal à $n - r$. En effet, la première des formules (7′) exprime la première inconnue principale x^1 sous forme de polynôme linéaire en inconnues paramétriques, les coefficients de ce polynôme étant les mêmes pour toutes les colonnes de la matrice X. Aussi la première ligne de la matrice est-elle une combinaison linéaire des $n - r$ dernières lignes. Autrement dit, si l'on multiplie les $n - r$ dernières lignes de la matrice X par les coefficients de la première des formules (7′) et qu'on additionne, on obtient la première ligne de cette matrice. De façon analogue à l'aide des autres formules (7′), on peut montrer que les lignes de numéros 2, ..., r sont des combinaisons linéaires des dernières lignes. Donc, par des transformations élémentaires, on peut rendre nulles les r premières lignes de la matrice X, ce qui signifie que son rang ne dépasse pas $n - r$.

On voit que le rang de la matrice considérée est $n - r$. Elle possède donc un mineur principal situé sur les colonnes $x_1, \ldots, x_{n-r}$, de sorte que la proposition 5 découle du théorème du mineur principal.

Soit $x_1, \ldots, x_{n-r}$ un ensemble fondamental de solutions d'un système d'équations linéaires homogènes. Considérons la matrice-colonne $x = C_1 x_1 + \ldots + C_{n-r} x_{n-r}$ et présentons-la sous la forme suivante :

$$\left\| \begin{matrix} x^1 \\ \ldots \\ x^n \end{matrix} \right\| = C_1 \left\| \begin{matrix} x_1^1 \\ \ldots \\ x_1^n \end{matrix} \right\| + \ldots + C_{n-r} \left\| \begin{matrix} x_{n-r}^1 \\ \ldots \\ x_{n-r}^n \end{matrix} \right\| . \qquad (13)$$

Il découle de la proposition 1 que la combinaison linéaire d'un ensemble quelconque de solutions du système homogène est aussi sa solution. Par conséquent, quels que soient les nombres $C_1, \ldots, C_{n-r}$, la matrice-colonne x définie par la formule (13) est une solution du système d'équations homogènes étudié. Inversement, selon la proposition 5, pour toute solution du système homogène il existe des nombres $C_1, \ldots, C_{n-r}$ pour lesquels cette solution prend la forme (13).

Une question se pose tout naturellement : est-ce qu'il existe un ensemble de $s < n - r$ solutions telles que chacune des solutions du système d'équations linéaires homogènes est une combinaison linéaire des solutions de l'ensemble considéré ? On démontre aisément que la réponse est négative. En effet, supposons qu'un tel ensemble de solutions existe. Considérons une matrice G de type $(n, n - r + s)$ dont les colonnes sont les solutions appartenant à cet ensemble et au système fondamental normal de solutions. D'une part, on a $\mathrm{Rg}\, G = s$ et, d'autre part, $\mathrm{Rg}\, G = n - r$. On aboutit donc à une contradiction.

5. Solution générale d'un système d'équations linéaires. On peut maintenant généraliser les résultats acquis dans les propositions 2 et 5.

THÉORÈME 2. *Si y_0 est une solution particulière du système* (1) *et $x_1, \ldots, x_{n-r}$ un ensemble fondamental de solutions du système homogène asso-*

cié à (1), *la matrice-colonne*

$$y = y_0 + C_1 x_1 + \ldots + C_{n-r} x_{n-r},$$

no bien écrite sous une forme développée,

$$\left\| \begin{matrix} y^1 \\ \ldots \\ y^n \end{matrix} \right\| = \left\| \begin{matrix} y_0^1 \\ \ldots \\ y_0^n \end{matrix} \right\| + C_1 \left\| \begin{matrix} x_1^1 \\ \ldots \\ x_1^n \end{matrix} \right\| + \ldots + C_{n-r} \left\| \begin{matrix} x_{n-r}^1 \\ \ldots \\ x_{n-r}^n \end{matrix} \right\|, \quad (14)$$

est solution du système d'équations linéaires (1) *pour tous nombres* $C_1, \ldots$ $\ldots, C_{n-r}$. *Inversement, pour toute solution de ce système il existe des nombres* $C_1, \ldots, C_{n-r}$ *pour lesquels elle prend la forme* (14).

L'expression se trouvant dans le second membre de l'égalité (14) est appelée *solution générale du système d'équations linéaires* (1).

Le théorème est vrai pour tout système d'équations linéaires et, en particulier, pour les systèmes homogènes. La formule (14) devient équivalente à (13) si y_0 est une solution triviale.

La règle de Cramer affirme que pour l'existence et l'unicité de la solution d'un système de n équations à n inconnues il suffit que le déterminant de la matrice du système soit différent de zéro. Le théorème 2 entraîne que cette condition est aussi nécessaire.

PROPOSITION 6. *Soit A la matrice du système de n équations linéaires à n inconnues. Si* $\det A = 0$, *le système soit n'a pas de solutions, soit en a une infinité.*

DÉMONSTRATION. L'égalité $\det A = 0$ signifie que le rang de la matrice A est strictement inférieur à n et, par suite, le système homogène associé a une infinité de solutions. Si le système considéré est compatible, il ressort de la formule (14) qu'il possède une infinité de solutions.

6. Exemples. Soit

$$Ax + By + Cz + D = 0 \quad (15)$$

l'équation du plan rapporté à un repère cartésien. Traitons-la comme un système d'une seule équation linéaire à trois inconnues. Posons $A \neq 0$, c'est donc un mineur principal de la matrice du système. Le rang de la matrice complète est au plus égal à l'unité, si bien que le système est compatible. On peut trouver l'une de ses solutions en posant $y = z = 0$. On obtient $x = -D/A$. Vu que $n = 3$ et $r = 1$, le système fondamental des solutions du système homogène associé comprend deux solutions. On les obtient en donnant aux inconnues paramétriques deux couples de valeurs : $y = 1, z = 0$ et $y = 0, z = 1$. Les valeurs correspondantes de l'inconnue principale x sont alors $-B/A$ et $-C/A$. Ainsi donc, la solution générale

du système (15) est

$$\left\|\begin{matrix} x \\ y \\ z \end{matrix}\right\| = \left\|\begin{matrix} -D/A \\ 0 \\ 0 \end{matrix}\right\| + C_1 \left\|\begin{matrix} -B/A \\ 1 \\ 0 \end{matrix}\right\| + C_2 \left\|\begin{matrix} -C/A \\ 0 \\ 1 \end{matrix}\right\| . \tag{16}$$

Dégageons le sens géométrique de la solution obtenue. Il est d'abord évident que la solution $(-D/A, 0, 0)$ représente les coordonnées d'un point (initial) du plan ou, ce qui revient au même, les composantes de son rayon vecteur. Dans la formule (14), y_0 est une solution quelconque du système, ce qui correspond au fait que le point initial peut être choisi d'une façon arbitraire. Selon la proposition 2 du § 2, ch. II, les composantes de tout vecteur situé dans le plan vérifient l'équation $A\alpha_1 + B\alpha_2 + C\alpha_3 = 0$, autrement dit le système homogène associé. A deux solutions linéairement indépendantes de ce système on peut faire correspondre deux vecteurs directeurs du plan. Ainsi, la formule (16) n'est autre chose que l'équation paramétrique du plan, où C_1 et C_2 sont des paramètres. L'arbitraire de choix du système fondamental de solutions se traduit donc par l'arbitraire de choix des vecteurs directeurs du plan.

A titre d'un autre exemple, considérons une droite dans l'espace. Rapportée au repère cartésien, elle peut être définie par un système d'équations

$$\left.\begin{aligned} A_1x + B_1y + C_1z + D_1 &= 0, \\ A_2x + B_2y + C_2z + D_2 &= 0. \end{aligned}\right\} \tag{17}$$

Admettons que le mineur $A_1B_2 - A_2B_1$ est différent de zéro et qu'il est donc principal. Vu que le rang de la matrice complète ne peut dépasser 2, le système est compatible.

Pour exprimer les inconnues principales x et y en fonction de l'inconnue paramétrique z, il est naturel dans notre cas (le système de deux équations à coefficients littéraux) d'utiliser non pas la méthode de Gauss mais la règle de Cramer.

Pour trouver une solution particulière du système, c'est-à-dire les composantes du rayon vecteur du point initial de la droite, posons $z_0 = 0$. Il vient

$$x_0 = \frac{B_1D_2 - B_2D_1}{A_1B_2 - A_2B_1}, \qquad y_0 = \frac{D_1A_2 - D_2A_1}{A_1B_2 - A_2B_1} .$$

Le système homogène associé prend la forme

$$\left.\begin{aligned} A_1x + B_1y + C_1z &= 0, \\ A_2x + B_2y + C_2z &= 0. \end{aligned}\right\} \tag{18}$$

Son ensemble fondamental de solutions comprend une seule solution. En

donnant à l'inconnue paramétrique z la valeur $z_1 = 1$, on obtient de (18)

$$x_1 = \frac{B_1C_2 - B_2C_1}{A_1B_2 - A_2B_1}, \qquad y_1 = \frac{C_1A_2 - C_2A_1}{A_1B_2 - A_2B_1}.$$

Toutefois il n'est plus commode de considérer ici l'ensemble fondamental normal de solutions. Multiplions donc la solution trouvée du système homogène par $A_1B_2 - A_2B_1$. Ainsi, la solution générale du système (17) est de la forme

$$\left\| \begin{matrix} x \\ y \\ z \end{matrix} \right\| = \left\| \begin{matrix} x_0 \\ y_0 \\ 0 \end{matrix} \right\| + C \left\| \begin{matrix} B_1C_2 - B_2C_1 \\ C_1A_2 - C_2A_1 \\ A_1B_2 - A_2B_1 \end{matrix} \right\|. \tag{19}$$

La même forme du vecteur directeur de la droite a été obtenue dans la proposition 11 du § 2, ch. II.

§ 6. Multiplication des matrices

1. Définition et exemples. Considérons d'abord une matrice-ligne $\boldsymbol{a}$ d'éléments a_i $(i = 1, \ldots, n)$ et une matrice-colonne $\boldsymbol{b}$ d'éléments b_j $(j = 1, \ldots, n)$. Il est essentiel que dans $\boldsymbol{a}$ et $\boldsymbol{b}$ le nombre d'éléments soit le même. On appelle *produit de* $\boldsymbol{a}$ *par* $\boldsymbol{b}$ la somme des produits d'éléments de même numéro, c'est-à-dire

$$ab = a_1b_1 + a_2b_2 + \ldots + a_nb_n.$$

Soient maintenant une matrice A à m lignes et n colonnes et une matrice B à n lignes et p colonnes. Les matrices sont telles que le nombre de colonnes de la première est égal au nombre de lignes de la seconde. Multiplions chaque ligne de A par chaque colonne de B. Ecrivons les mp produits obtenus sous forme de matrice C à m lignes et p colonnes. Plus précisément, les éléments de chaque colonne de la matrice C sont les produits de toutes les lignes de la matrice A par la colonne correspondante de la matrice B, et les éléments de chaque ligne de C sont les produits de la ligne correspondante de A par toutes les colonnes de B. Ainsi, les éléments de la matrice C peuvent être représentés par la formule :

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} \tag{1}$$

$$(i = 1, \ldots, m\ ; \quad j = 1, \ldots, p).$$

DÉFINITION. On appelle *produit d'une matrice A par une matrice B* et on note AB la matrice C dont les éléments sont exprimés en fonctions des éléments de A et B par les formules (1).

La définition du produit des matrices est plus compliquée et paraît moins naturelle que celle de la somme. Cependant, si on essayait de définir le produit des matrices de mêmes dimensions comme la matrice obtenue en multipliant leurs éléments homologues, une telle définition ne trouverait guère d'applications importantes. Quant à la définition donnée, elle est, comme le lecteur s'en convaincra, largement utilisée.

Considérons quelques exemples.

1) Le produit d'une matrice carrée A d'ordre n par une matrice-colonne x à n éléments :

$$Ax = \begin{Vmatrix} a_1^1 \dots a_n^1 \\ a_1^2 \dots a_n^2 \\ \dots\dots\dots \\ a_1^n \dots a_n^n \end{Vmatrix} \begin{Vmatrix} x^1 \\ x^2 \\ \dots \\ x^n \end{Vmatrix} = \begin{Vmatrix} a_1^1 x^1 + \dots + a_n^1 x^n \\ a_1^2 x^1 + \dots + a_n^2 x^n \\ \dots\dots\dots\dots\dots \\ a_1^n x^1 + \dots + a_n^n x^n \end{Vmatrix}.$$

Le produit est une matrice-colonne à n éléments. On ne peut multiplier les matrices dans l'ordre inverse, autrement dit xA n'est pas défini.

2) Le produit d'une matrice-ligne à n-éléments par une matrice carrée d'ordre n est une matrice-ligne à n éléments :

$$\| x_1 \dots x_n \| \begin{Vmatrix} a_{11} \dots a_{1n} \\ a_{21} \dots a_{2n} \\ \dots\dots\dots \\ a_{n1} \dots a_{nn} \end{Vmatrix} = \left\| \sum_{k=1}^{n} a_{k1} x_k \dots \sum_{k=1}^{n} a_{kn} x_k \right\|.$$

3) Le produit d'une matrice-colonne à m éléments par une matrice-ligne à n éléments est une matrice de type (m, n) :

$$\begin{Vmatrix} x^1 \\ \dots \\ x^m \end{Vmatrix} \| a_1 \dots a_n \| = \begin{Vmatrix} x^1 a_1 & x^1 a_2 & \dots & x^1 a_n \\ x^2 a_1 & x^2 a_2 & \dots & x^2 a_n \\ \dots & \dots & \dots & \dots \\ x^m a_1 & x^m a_2 & \dots & x^m a_n \end{Vmatrix}.$$

4) Soient A une matrice de type (m, n), e_i la i-ième colonne de la matrice unité *) d'ordre m et e_j la j-ième colonne de la matrice unité d'ordre n. Alors

$${}^t e_i A e_j = \| 0 \dots 1 \dots 0 \| \begin{Vmatrix} a_{11} \dots a_{1n} \\ \dots\dots\dots \\ a_{m1} \dots a_{mn} \end{Vmatrix} \begin{Vmatrix} 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{Vmatrix} = a_{ij}.$$

---

*) Voir la définition de la matrice unité à la p. 131.

PROPOSITION 1. *La j-ième colonne de la matrice AB est une combinaison linéaire des colonnes de A dont les coefficients sont égaux aux éléments de la j-ième colonne de la matrice B.*

La i-ième ligne de la matrice AB est une combinaison linéaire des lignes de B dont les coefficients sont égaux aux éléments de la i-ième ligne de la matrice A.

Les deux assertions se démontrent de la même façon. Démontrons par exemple la première. Pour le faire, désignons les colonnes des matrices A, B et AB respectivement par $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ et $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_p$. Remarquons que les colonnes des matrices A et AB ont le même nombre d'éléments. Etant donné que pour obtenir $\boldsymbol{c}_j$ on multiplie successivement toutes les lignes de A par $\boldsymbol{b}_j$, on a $\boldsymbol{c}_j = A\boldsymbol{b}_j$. Il ressort de l'exemple 1) que le produit $A\boldsymbol{b}_j$ peut être écrit : $b_{j1}\boldsymbol{a}_1 + \ldots + b_{jn}\boldsymbol{a}_n$, ce qui démontre la proposition.

2. Propriétés de la multiplication des matrices. La multiplication des matrices n'est pas commutative. Même si les deux produits AB et BA sont définis, ils peuvent ne pas être égaux comme le montre l'exemple suivant :

$$\begin{Vmatrix} 1 & 1 \\ 0 & 0 \end{Vmatrix} \begin{Vmatrix} 0 & 0 \\ 1 & 1 \end{Vmatrix} = \begin{Vmatrix} 1 & 1 \\ 0 & 0 \end{Vmatrix} \neq \begin{Vmatrix} 0 & 0 \\ 1 & 1 \end{Vmatrix} = \begin{Vmatrix} 0 & 0 \\ 1 & 1 \end{Vmatrix} \begin{Vmatrix} 1 & 1 \\ 0 & 0 \end{Vmatrix}.$$

Si deux matrices A et B vérifient la relation $AB = BA$, on dit qu'elles commutent ou sont *commutables*. De telles matrices existent. Par exemple la matrice unité d'ordre n

$$E_n = \begin{Vmatrix} 1 & 0 & 0 & \ldots & 0 & 0 \\ 0 & 1 & 0 & \ldots & 0 & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{Vmatrix}$$

commute avec toute matrice carrée de même ordre, c'est-à-dire

$$AE_n = E_nA = A. \tag{2}$$

Laissons au lecteur le soin de vérifier ces égalités à titre d'exercice sur la multiplication des matrices.

Le fait qu'une matrice A ne change pas par multiplication par la matrice unité est une propriété importante de la matrice unité à laquelle elle doit son nom. Si une autre matrice $\tilde{E}$ possédait la même propriété, on aurait $\tilde{E}E = E$ et $\tilde{E}E = \tilde{E}$, d'où $E = \tilde{E}$.

Il va de soi que, quelles que soient les matrices A et B, on a

$$AO = O, \qquad OB = O,$$

où O est la matrice nulle. (On admet que les produits sont définis.)

PROPOSITION 2. *La multiplication des matrices est associative, c'est-à-*

dire si les produits AB et $(AB)C$ sont définis, il en est de même de BC et de $A(BC)$, de sorte qu'on a l'égalité $(AB)C = A(BC)$.

En effet, soient les matrices A, B et C de types respectifs (m_A, n_A), (m_B, n_B) et (m_C, n_C). Si AB est défini, $n_A = m_B$ et la matrice AB est de type (m_A, n_B). Donc, si le produit $(AB)C$ est défini, $n_B = m_C$. Les éléments de la matrice AB sont

$$\sum_{\alpha=1}^{n_A} a_{i\alpha} b_{\alpha\rho} \qquad (i = 1, \ldots, m_A\ ;\ \rho = 1, \ldots, n_B)$$

et, par suite, ceux de $(AB)C$ sont de la forme

$$\sum_{\rho=1}^{n_B} \left(\sum_{\alpha=1}^{n_A} a_{i\alpha} b_{\alpha\rho} \right) c_{\rho\lambda} \quad (i = 1, \ldots, m_A\ ;\ \lambda = 1, \ldots, n_C). \tag{3}$$

Puisque $n_B = m_C$, le produit BC est défini. Ses éléments sont

$$\sum_{\rho=1}^{n_B} b_{\alpha\rho} c_{\rho\lambda} \qquad (\alpha = 1, \ldots, m_A\ ;\ \lambda = 1, \ldots, n_C).$$

Le nombre de lignes de la matrice BC est égal à n_A, autrement dit au nombre de colonnes de la matrice A. Ainsi est défini le produit $A(BC)$. Les éléments de $A(BC)$ sont de la forme

$$\sum_{\alpha=1}^{n_A} a_{i\alpha} \left(\sum_{\rho=1}^{n_B} b_{\alpha\rho} c_{\rho\lambda} \right) \quad (i = 1, \ldots, m_A\ ;\ \lambda = 1, \ldots, n_C). \tag{4}$$

En vertu des propositions 1 et 2 du § 2, les expressions (3) et (4) coïncident, si bien que la proposition 2 est démontrée.

PROPOSITION 3. *La multiplication des matrices est distributive par rapport à l'addition, c'est-à-dire si $A(B + C)$ est définie, on a*

$$A(B + C) = AB + AC.$$

Si $(A + B)C$ est définie, on a

$$(A + B)C = AC + BC.$$

Les deux parties de la proposition se démontrent de la même façon. Démontrons par exemple la première. Il est évident que B et C doivent être de type (m, n), et A de type (p, m) (p peut être quelconque). On peut exprimer les éléments de la matrice $A(B + C)$ en fonction des éléments des matrices A, B et C :

$$\sum_{i=1}^{m} a_{\lambda i}(b_{i\alpha} + c_{i\alpha}) \qquad (\lambda = 1, \ldots, p\ ;\ \alpha = 1, \ldots, n).$$

Selon la proposition 1 du § 2, cette somme peut être représentée sous la forme

$$\sum_{i=1}^{m} a_{\lambda i} b_{i\alpha} + \sum_{i=1}^{m} a_{\lambda i} c_{i\alpha}.$$

Les sommes qui composent cette expression sont égales aux éléments des matrices AB et AC, situés à l'intersection de la ligne de numéro λ et de la colonne de numéro α. La proposition est ainsi démontrée.

On démontre aisément la propriété suivante de la multiplication des matrices.

PROPOSITION 4. *Si le produit AB est défini, on a*

$$\alpha(AB) = (\alpha A)B = A(\alpha B)$$

pour tout nombre α.

PROPOSITION 5. *Le rang du produit de deux matrices ne dépasse pas les rangs des facteurs.*

Pour le démontrer, considérons une matrice D composée de toutes les colonnes des matrices A et AB. Ecrivons-la sous la forme : $D = \| A \mid AB \|$. Il est évident que $\operatorname{Rg} AB \leqslant \operatorname{Rg} D$. Selon la proposition 1, les colonnes de AB sont des combinaisons linéaires des colonnes de A. D'où il ressort que $\operatorname{Rg} D = \operatorname{Rg} A$. On a $\operatorname{Rg} AB \leqslant \operatorname{Rg} A$. De façon analogue on démontre que $\operatorname{Rg} AB \leqslant \operatorname{Rg} B$. Il faut pour cela considérer la matrice D' composée des lignes de la matrice B et des lignes de la matrice AB.

PROPOSITION 6. *Si le produit AB est défini, il en est de même du produit ${}^tB\,{}^tA$ et on a l'égalité*

$${}^t(AB) = {}^tB\,{}^tA.$$

Supposons que les matrices A et B sont respectivement de types (m, n) et (n, p). L'élément de la matrice AB situé à l'intersection de la i-ième ligne et de la j-ième colonne est alors de la forme

$$\sum_{\alpha=1}^{n} a_{i\alpha} b_{\alpha j} \qquad (i = 1, \dots, m\ ; \quad j = 1, \dots, p). \tag{5}$$

La j-ième ligne de la matrice ${}^tB$ est composée des éléments $b_{1j}, \dots, b_{nj}$, tandis que la i-ième colonne de la matrice ${}^tA$ des éléments $a_{i1}, \dots, a_{in}$. Donc, le produit ${}^tB\,{}^tA$ est défini, et l'élément situé à l'intersection de la j-ième ligne et de la i-ième colonne est de la forme

$$\sum_{\alpha=1}^{n} b_{\alpha j} a_{i\alpha}.$$

Il se confond avec l'élément (5), les indices i et j parcourant dans les deux

expressions les mêmes valeurs $i = 1, \ldots, m$; $j = 1, \ldots, p$. La proposition est ainsi démontrée.

Corollaire. *Si le produit ABC est défini, on a*

$$^t(ABC) = {}^tC\,{}^tB\,{}^tA.$$

En effet, $^t(ABC) = {}^tC\,{}^t(AB) = {}^tC\,{}^tB\,{}^tA$.

3. Matrice inverse. Etudions maintenant les possibilités d'introduire l'opération de division qui est l'inverse de celle de multiplication.

Définition. Etant donné une matrice A, on dit que X est la *matrice inverse* de A, et on la note A^{-1}, si

$$XA = AX = E_n, \tag{6}$$

où E_n est la matrice unité d'ordre n.

Puisque A et A^{-1} commutent, elles doivent toutes les deux être carrées et de même ordre n. En effet, il ressort de (6), en vertu de la proposition 5, que $\mathrm{Rg}\,E_n \leqslant \mathrm{Rg}\,A$. D'où $\mathrm{Rg}\,A = n$. Par suite, la matrice A peut avoir une inverse si son déterminant n'est pas nul. La condition mentionnée est non seulement nécessaire mais aussi suffisante pour que la matrice inverse existe.

Proposition 7. *Toute matrice carrée de déterminant non nul admet une matrice inverse et une seule.*

Démonstration. Pour toute matrice A, avec $\det A \neq 0$, il existe une matrice X et une seule telle que $AX = E$. En effet, la colonne x_j de la matrice X doit satisfaire pour tout j à la condition $Ax_j = e_j$, où e_j est la j-ième colonne de la matrice unité. De façon plus détaillée, cette condition s'écrit par un système d'équations linéaires

$$\left.\begin{array}{r} a_{11}x_j^1 + \ldots + a_{1n}x_j^n = 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{j1}x_j^1 + \ldots + a_{jn}x_j^n = 1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{n1}x_j^1 + \ldots + a_{nn}x_j^n = 0. \end{array}\right\} \tag{7}$$

D'après la règle de Cramer, ce système d'équations admet une solution unique et, par suite, chaque colonne de la matrice X est parfaitement définie.

Démontrons que $XA = E$. A cet effet remarquons que $\det X \neq 0$, de sorte que d'après ce qui vient d'être démontré il existe une matrice Y telle que $XY = E$. On obtient Y en multipliant à gauche les deux membres de la dernière égalité par la matrice A. Il vient $AXY = A$, d'où, en vertu de $AX = E$, on a $Y = A$. Ainsi donc, la matrice X satisfait aux deux conditions (6).

Le procédé utilisé dans la démonstration de l'existence sert de base dans la recherche de la matrice inverse. Selon la règle de Cramer, on obtient la i-ième inconnue du système (7) à partir de la formule $x_j^i = \Delta_i/\det A$, où Δ_i est le déterminant de la matrice déduite de A par substitution à sa i-ième colonne de la j-ième colonne de la matrice unité. En développant Δ_i suivant cette colonne, on obtient un seul terme car un seul élément dans e_j est égal à 1, les autres étant des zéros. Par conséquent, $\Delta_i = (-1)^{i+j}M_i^j$, où M_i^j est le mineur associé à l'élément a_i^j dans la matrice A. En définitive :

$$x_j^i = \frac{(-1)^{i+j}M_i^j}{\det A}. \tag{8}$$

Le système (7) peut également être résolu par la méthode de Gauss. Puisque $\det A \neq 0$, le corollaire de la proposition 3 du § 4 nous dit qu'il est possible de transformer la matrice A en matrice unité par les opérations élémentaires sur les lignes de la matrice complète du système. Ceci étant, la matrice-colonne des termes constants e_j devient la solution x_j du système (comp. les formules (7) du § 5, qui dans le cas considéré ne contiennent pas d'inconnues paramétriques).

Pour les j différents, les systèmes (7) ne se distinguent que par les matrices-colonnes des termes constants. Aussi les opérations élémentaires sur les lignes de la matrice complète sont-elles les mêmes pour tous les j. On peut résoudre tous les systèmes à la fois en joignant à A toutes les colonnes des termes constants des systèmes :

$$\left\| \begin{array}{ccc|ccccc} a_{11} & \dots & a_{1n} & 1 & 0 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & \dots & a_{nn} & 0 & 0 & \dots & 0 & 1 \end{array} \right\|. \tag{9}$$

On obtient de ce qui vient d'être dit que si les opérations élémentaires sur les lignes de la matrice (9) transforment sa partie gauche en matrice unité, sa partie droite devient la matrice A^{-1}.

Citons quelques propriétés de la matrice inverse : il découle immédiatement de la définition que

$$(A^{-1})^{-1} = A.$$

En outre,

$$(AB)^{-1} = B^{-1}A^{-1}, \tag{10}$$

$$({}^tA)^{-1} = {}^t(A^{-1}). \tag{11}$$

Vérifions la formule (10) :

$$AB(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = E.$$

Pour démontrer la formule (11), transposons les deux membres de

l'égalité $AA^{-1} = E$. On obtient ${}^t(A^{-1})^tA = E$, d'où se déduit la formule (11).

4. Transformations élémentaires en tant que multiplication des matrices. Déterminant du produit. Toute transformation élémentaire de lignes d'une matrice A à m lignes et n colonnes est équivalente à la multiplication à gauche de A par une matrice carrée d'ordre m. Plus précisément, considérons la matrice S_1 déduite de la matrice unité par permutation des lignes i et j. Il ressort immédiatement de la proposition 1 que la multiplication à gauche de A par S_1 fait changer l'ordre des lignes correspondantes dans la matrice A.

Soit S_2 la matrice déduite de la matrice unité par substitution de $\lambda \neq 0$ à 1 situé à la i-ième position sur la diagonale. Il découle de la proposition 1 que la i-ième ligne de la matrice obtenue par multiplication à gauche de A par S_2 se trouve multipliée par λ.

Notons S_3 la matrice obtenue de la matrice unité en substituant l'unité à l'élément nul situé à l'intersection de la i-ième ligne et de la j-ième colonne. La multiplication à gauche de A par S_3 s'avère équivalente à l'addition de la j-ième ligne de A à la i-ième.

Signalons que les déterminants de S_1, S_2 et S_3 sont différents de zéro : $\det S_1 = -1$, $\det S_2 = \lambda$, $\det S_3 = 1$. Si A est une matrice carrée, on a

$$\det(S_1A) = -\det A,\quad \det(S_2A) = \lambda \det A,\quad \det(S_3A) = \det A.$$

Ainsi, pour les matrices de transformations élémentaires,

$$\det(SA) = \det S \det A. \tag{12}$$

Aux transformations élémentaires successives correspond la multiplication à gauche par le produit des facteurs matriciels correspondants.

On peut réaliser les transformations élémentaires des colonnes de A par multiplications à droite de A par des matrices analogues.

PROPOSITION 8. *Si* $\det A \neq 0$, *il existe des matrices de transformations élémentaires* $S_1, \ldots, S_p$ *telles que* $A = S_1 \ldots S_p$.

DÉMONSTRATION. Si $\det A \neq 0$, il existe A^{-1}. Vu que $\det A^{-1} \neq 0$, les opérations élémentaires sur les lignes de la matrice A^{-1} permettent de la transformer en matrice unité. Il existe donc des matrices de transformations élémentaires $S_1, \ldots, S_p$ pour lesquelles $S_1 \ldots S_pA^{-1} = E$. Ce sont évidemment les matrices cherchées.

On est maintenant en mesure de démontrer la

PROPOSITION 9. *Pour toutes matrices carrées A et B du même ordre,* $\det(AB) = \det A \det B$.

En effet, si $\det A = 0$, l'assertion découle de la proposition 5 sur le rang du produit des matrices. Mais si $\det A \neq 0$, la matrice A peut être pré-

sentée sous forme de produit des matrices de transformations élémentaires : $A = S_1 \dots S_p$. En appliquant maintenant la formule (12), on obtient

$$\det(AB) = \det(S_1 \dots S_p B) = \det S_1 \dots \det S_p \det B = \det A \det B.$$

5. Matrices complexes. Les propriétés des quatre opérations arithmétiques étant les mêmes pour les nombres réels aussi bien que pour les nombres complexes, toutes les assertions sur les nombres réels où n'interviennent que ces propriétés sont aussi vraies pour les nombres complexes. En particulier, par analogie au § 2, on peut définir le déterminant de la matrice carrée dont les éléments sont des nombres complexes. De par la définition c'est un nombre complexe, et toutes les propositions démontrées au § 2 demeurent vraies.

En étudiant les systèmes d'équations linéaires dans les §§ 3, 4 et 5 on n'a utilisé que les propriétés de l'addition, de la soustraction, de la multiplication et de la division, qui sont les mêmes pour les nombres réels et complexes. Donc, tout ce qui a été démontré s'applique aussi aux systèmes d'équations linéaires à coefficients et seconds membres complexes.

Les matrices à éléments complexes vérifient tout ce qui a été dit sur les matrices dans les §§ 1 et 6. La définition et les propriétés de l'addition et de la multiplication des matrices, y compris la construction de la matrice inverse, sont formulées et démontrées de façon identique pour les nombres réels et complexes.

A proprement parlé, dans les paragraphes précédents on n'a pas spécifié la nature des nombres, ce qui d'ailleurs n'était pas nécessaire vu que tout ce qui a été dit se rapportait dans la même mesure aux nombres réels et complexes. Soulignons toutefois que dans les combinaisons linéaires de matrices-colonnes à éléments complexes on peut prendre pour coefficients des nombres complexes quelconques.

Arrêtons-nous sur quelques particularités des matrices complexes. Soient z_{kj} les éléments de la matrice Z. Si $z_{kj} = a_{kj} + b_{kj}i$, la matrice Z peut être représentée sous la forme

$$Z = A + Bi,$$

où les matrices A et B sont composées d'éléments a_{kj} et b_{kj} respectivement. On appelle A *partie réelle* et B *partie imaginaire* de la matrice Z. Deux matrices complexes Z et Z_1 sont égales si et seulement si $A = A_1$ et $B = B_1$. Ainsi, à une égalité de deux matrices complexes correspondent deux égalités entre les matrices réelles. On remarque ici une coïncidence parfaite avec l'égalité des nombres complexes.

La matrice composée d'éléments $\bar{z}_{kj}$ est appelée *conjuguée* de la matrice Z et est notée $\bar{Z}$. Pour toutes les matrices complexes on a

$$\overline{Z_1 + Z_2} = \bar{Z}_1 + \bar{Z}_2, \qquad (13)$$

$$\overline{Z_1 Z_2} = \bar{Z}_1 \bar{Z}_2, \tag{14}$$

$$\overline{\det Z} = \det \bar{Z} \tag{15}$$

(à condition que la somme, le produit et le déterminant qui figurent dans l'un des membres de l'égalité soient définis).

En démontrant la dernière relation, on peut se servir du développement (6) du § 2. Il vient

$$\overline{\det Z} = \overline{\sum_{(j_1 \ldots j_n)} (-1)^{N(j_1, \ldots, j_n)} z_{1j_1} \ldots z_{nj_n}} =$$

$$= \sum_{(j_1 \ldots j_n)} (-1)^{N(j_1, \ldots, j_n)} \bar{z}_{1j_1} \ldots \bar{z}_{nj_n} = \det \bar{Z}.$$

Les égalités (13) et (14) se démontrent de la même façon.

Une matrice est réelle (c'est-à-dire est composée de nombres réels) si et seulement si elle est égale à sa matrice conjuguée. En effet, chacun de ses éléments est égal dans ce cas à son conjugué.

ÇHAPITRE VI

ESPACES VECTORIELS

§ 1. Notions générales

1. Définition d'un espace vectoriel. Dans ce livre, on a déjà rencontré des ensembles munis des opérations d'addition et de multiplication par un nombre. Dans le chapitre I par exemple, on a considéré l'ensemble des vecteurs (segments orientés) dans lequel à tout couple de vecteurs on a associé, suivant la règle du parallélogramme, un vecteur appelé leur somme, et à tout vecteur **a** et tout nombre α, un vecteur appelé produit de **a** par α.

Dans l'ensemble des matrices de mêmes dimensions, on a introduit l'opération d'addition en appelant somme de deux matrices la matrice dont les éléments sont égaux aux sommes des éléments homologues des termes. On a de même introduit l'opération de multiplication d'une matrice par un nombre : ce produit est la matrice dont les éléments sont les produits des éléments de la matrice donnée par ce nombre. Si les éléments de la matrice sont des nombres complexes, on peut aussi définir le produit de cette matrice par un nombre complexe. On a vu que les propriétés de ces opérations, exprimées dans la proposition 1 du § 1, ch. V, étaient les mêmes que celles des opérations sur les vecteurs formulées dans la proposition 1 du § 1, ch. I.

Bien que définies de façon différente dans chacun de ces ensembles les opérations considérées ont des propriétés communes : commutativité et associativité de l'addition, distributivité de la multiplication par un nombre relativement à l'addition des nombres, etc. Dans les calculs des vecteurs ou des matrices, on n'utilise que ces propriétés sans faire intervenir les définitions des opérations. On donnera plus loin d'autres exemples d'ensembles où sont définies des opérations possédant les mêmes propriétés.

Il devient, tout naturellement, nécessaire d'étudier un ensemble de nature quelconque où sont définies les opérations d'addition de deux éléments et de multiplication d'un élément par un nombre. Ces opérations peuvent être définies de façon quelconque à condition qu'elles possèdent une gamme de propriétés bien déterminée.

DÉFINITION. L'ensemble $\mathscr{L}$ sera appelé *espace vectoriel* et ses éléments *vecteurs* si :

1) est définie une loi de composition interne (*opération d'addition*) suivant laquelle à deux éléments quelconques x et y de $\mathscr{L}$ est associé un élément appelé leur *somme*, noté $x + y$;

2) est définie une loi de composition externe (*opération de multiplication par un nombre*) suivant laquelle à un élément x de $\mathscr{L}$ et à un nombre α est associé un élément de $\mathscr{L}$ appelé *produit* de x par α, noté αx ;

3) pour tous éléments x, y et z de $\mathscr{L}$ et tous nombres α et β, sont remplies les conditions (ou axiomes) suivantes :

1° $x + y = y + x$;

2° $(x + y) + z = x + (y + z)$;

3° il existe un élément o tel que pour tout x de $\mathscr{L}$ on a $x + o = x$;

4° pour chaque x il existe un élément $-x$ tel que $x + (-x) = o$;

5° $\alpha(x + y) = \alpha x + \alpha y$;

6° $(\alpha + \beta)x = \alpha x + \beta x$;

7° $\alpha(\beta x) = (\alpha\beta)x$;

8° le produit de tout élément x par le nombre 1 est égal à x, c'est-à-dire $1x = x$.

Si au point 2) on se limite aux nombres réels, $\mathscr{L}$ est appelé *espace vectoriel réel* ; si est définie la multiplication par tout nombre complexe, l'espace vectoriel $\mathscr{L}$ est dit *complexe*.

Le vecteur $-x$ est dit *opposé* du vecteur x. Le vecteur o est appelé *vecteur nul* ou *zéro*.

On désignera les vecteurs par des lettres latines minuscules et les nombres par des lettres grecques.

Donnons encore quelques exemples d'espaces vectoriels.

EXEMPLE 1. Considérons l'ensemble de toutes les fonctions d'une variable indépendante ξ, définies et continues pour $0 \leqslant \xi \leqslant 1$. A deux fonctions quelconques $f(\xi)$ et $g(\xi)$ de cet ensemble on peut associer leur somme $f(\xi) + g(\xi)$ qui est aussi une fonction définie et continue pour $0 \leqslant \xi \leqslant 1$ et qui, par suite, appartient à l'ensemble considéré. A un nombre α et à une fonction $f(\xi)$ on peut associer une fonction $\alpha f(\xi)$ (produit banal d'une fonction par un nombre) qui est définie et continue pour $0 \leqslant \xi \leqslant 1$ si la fonction $f(\xi)$ l'est. Les huit axiomes sont vérifiés. Le rôle de zéro est joué par la fonction identiquement nulle.

EXEMPLE 2. Soit $\mathscr{L}$ l'ensemble de tous les polynômes d'une seule variable, dont le degré est au plus égal à un nombre donné n. La somme de deux polynômes de $\mathscr{L}$ est également un polynôme de degré au plus égal à n ; de même, le produit d'un polynôme de $\mathscr{L}$ par un nombre appartient à $\mathscr{L}$. On vérifie facilement que les axiomes de l'espace vectoriel sont aussi vrais dans

ce cas. Le rôle de zéro est joué par le polynôme dont tous les coefficients sont nuls. $\mathcal{P}$ est un espace vectoriel réel ou complexe suivant que les coefficients des polynômes sont réels ou complexes.

EXEMPLE 3. L'ensemble des nombres complexes muni des opérations d'addition et de multiplication par un nombre complexe est un espace vectoriel complexe. D'une façon analogue, l'ensemble des nombres réels muni des opérations ordinaires d'addition et de multiplication est un espace vectoriel réel.

EXEMPLE 4. L'ensemble des nombres complexes muni des opérations ordinaires d'addition et de multiplication par un nombre réel constitue un espace vectoriel réel.

EXEMPLE 5. Il existe un espace vectoriel composé d'un seul élément. Cet espace est dit *nul*. L'élément unique y joue à la fois le rôle de zéro et de son propre opposé. Les opérations sont définies par les égalités $o + o = o$ et $\alpha o = o$.

2. Corollaires immédiats. Il ressort des axiomes de l'espace vectoriel qu'il existe un vecteur nul et un seul et que chaque vecteur admet un vecteur opposé et un seul. En effet, supposons qu'il existe deux vecteurs o_1 et o_2 vérifiant l'axiome 3°. Leur somme doit alors être égale à chacun d'eux : $o_1 + o_2 = o_1 = o_2$. D'une façon analogue, si un vecteur x possède deux vecteurs opposés $-x_1$ et $-x_2$, la somme $(-x_1) + x + (-x_2)$ doit être égale à $-x_1$ et $-x_2$.

L'égalité $o + o = o$ veut dire que l'opposé du vecteur nul est le vecteur nul, tandis que l'égalité $(-x) + (x) = o$ signifie que l'opposé du vecteur $-x$ est le vecteur x.

La somme des vecteurs y et $-x$ sera notée $y - x$ et appelée *différence* des vecteurs y et x.

On voit aisément que tout vecteur x vérifie l'égalité $0x = o$. En effet,

$$0x = 0x + x - x = (1 + 0)x - x = o.$$

Il en découle que $(-1)x = -x$ pour tout x. En effet,

$$(-1)x + x = (1 - 1)x = 0x = o.$$

Signalons de même que le produit de tout nombre par le vecteur nul est égal au vecteur nul, puisque

$$\alpha o = \alpha(x - x) = \alpha x - \alpha x = o.$$

Si $\alpha x = o$, on a : ou bien $\alpha = 0$, ou bien $x = o$. En effet, soit $\alpha \neq 0$. On peut alors multiplier l'égalité donnée par α^{-1} et obtenir $1x = o$.

L'expression de la forme $\alpha_1 x_1 + \ldots + \alpha_n x_n$ sera appelée, comme dans les chapitres précédents, *combinaison linéaire* des vecteurs $x_1, \ldots, x_n$ avec coefficients $\alpha_1, \ldots, \alpha_n$.

De tout ce qui vient d'être dit il ressort que les opérations sur les vecteurs dans un espace vectoriel sont effectuées suivant les mêmes règles que les opérations sur les vecteurs (segments orientés) de l'espace géométrique ordinaire.

3. Dépendance linéaire. Par analogie avec les définitions respectives pour les vecteurs et matrices-colonnes introduites dans les chapitres I et V, on définit le système libre et le système lié de vecteurs dans un espace vectoriel. Rappelons qu'une combinaison linéaire est dite triviale si tous ses coefficients sont nuls.

Définition. On dit qu'un système de vecteurs est lié s'il existe une combinaison linéaire non triviale de ces vecteurs qui est égale à zéro. Dans le cas contraire, c'est-à-dire lorsque la seule combinaison linéaire triviale des vecteurs est nulle, on dit que le système de vecteurs est libre.

Toutes les propositions démontrées pour les systèmes libres et liés de matrices-colonnes sont également vraies pour les systèmes de vecteurs. On se limitera ici à l'énoncé de ces propositions, car leurs démonstrations sont identiques à celles des propositions sur les matrices-colonnes (voir propositions 2 à 5, § 1, ch. V).

Proposition 1. *Un système de $k > 1$ vecteurs est lié si et seulement si au moins un des vecteurs est une combinaison linéaire des autres.*

Proposition 2. *Tout système qui contient le vecteur nul est lié.*

Proposition 3. *Le système de vecteurs est lié s'il contient un ensemble de vecteurs linéairement dépendants.*

Proposition 4. *Tout système extrait d'un système de vecteurs libre est encore libre.*

4. Base. La définition suivante jouera un grand rôle dans l'exposé qui suit.

Définition. On dit que le système fini et ordonné de vecteurs est une *base* de l'espace $\mathscr{L}$ si : a) il est libre et b) tout vecteur de $\mathscr{L}$ est une combinaison linéaire des vecteurs de ce système.

Notons que la base est par définition un système ordonné de vecteurs. Cela signifie qu'à chaque vecteur est attaché un numéro. En changeant l'ordre des vecteurs, on peut obtenir des bases différentes à partir d'un même système.

Les coefficients de la combinaison linéaire dont il s'agit dans la définition s'appellent *composantes* ou *coordonnées* du vecteur dans la base donnée.

On écrira les vecteurs $e_1, \ldots, e_n$ de la base sous la forme d'une matrice-ligne : $e = \| e_1 \ldots e_n \|$ et les composantes $\xi^1, \ldots, \xi^n$ du vecteur x dans la base

e sous la forme d'une matrice-colonne :

$$\xi = \begin{Vmatrix} \xi^1 \\ \dots \\ \xi^n \end{Vmatrix},$$

qu'on appellera *colonne de coordonnées* du vecteur.

On peut maintenant écrire le développement du vecteur x suivant les vecteurs de base sous l'une des formes suivantes*)

$$x = \sum_{i=1}^{n} \xi^i e_i = \| e_1 \dots e_n \| \cdot \begin{Vmatrix} \xi^1 \\ \dots \\ \xi^n \end{Vmatrix} = e\xi.$$

PROPOSITION 5. *Etant donné une base, les composantes de tout vecteur sont définies de façon univoque.*

Dans le cas contraire, on aurait deux égalités $x = \sum \xi^i e_i$ et $x = \sum \xi_1^i e_i$, d'où $\sum (\xi^i - \xi_1^i) e_i = o$. Puisque les vecteurs $e_1, \dots, e_n$ sont linéairement indépendants, tous les coefficients de la combinaison linéaire sont nuls et, par suite, $\xi^i = \xi_1^i$ pour tous $i = 1, \dots, n$.

PROPOSITION 6. *La colonne de coordonnées de la somme de deux vecteurs est égale à la somme de leurs colonnes de coordonnées. La colonne de coordonnées du produit d'un vecteur par un nombre est égale au produit de la colonne de coordonnées du vecteur donné par ce nombre.*

Pour le démontrer, il suffit d'écrire les suites d'égalités :

$$x + y = e\xi + e\eta = e(\xi + \eta)$$

et

$$\alpha x = \alpha e \xi = e(\alpha \xi),$$

où ξ et η sont les colonnes de coordonnées des vecteurs x et y. On se sert ici des propriétés de multiplication des matrices (propositions 3 et 4 du § 6, ch. V).

Il ressort immédiatement de la proposition 6 que la colonne de coordonnées de la combinaison linéaire des vecteurs est une combinaison linéaire de leurs colonnes de coordonnées avec les mêmes coefficients. D'où il découle la

PROPOSITION 7. *Les vecteurs sont linéairement dépendants si et seulement si leurs colonnes de coordonnées sont linéairement dépendantes.*

La démonstration est évidente : si la combinaison linéaire non triviale

---

*) Les éléments de la matrice-ligne e sont des vecteurs et non des nombres. On peut étendre à ces matrices-lignes toutes les opérations sur les matrices.

des vecteurs est nulle, il en est de même de la combinaison linéaire, avec les mêmes coefficients, de leurs colonnes de coordonnées. La proposition réciproque se démontre de la même façon.

THÉORÈME 1. *Si dans un espace vectoriel il existe une base de n vecteurs, tout autre base de cet espace est aussi composé de n vecteurs.*

En effet, soient dans l'espace deux bases $\|e_1, \ldots, e_n\|$ et $\|f_1, \ldots, f_m\|$, avec $m > n$. Décomposons chacun des vecteurs $f_1, \ldots, f_m$ suivant les vecteurs $e_1, \ldots, e_n$ et construisons une matrice dont les colonnes sont les colonnes de coordonnées obtenues. On a m matrices-colonnes dont chacune a n éléments. La matrice ainsi obtenue a donc n lignes et m colonnes et son rang ne dépasse pas n. En vertu du théorème 2 du § 4, ch. V, les colonnes de la matrice sont linéairement dépendantes et, partant, sont dépendants les vecteurs $f_1, \ldots, f_m$: contradiction.

On est maintenant en droit d'introduire la définition suivante.

DÉFINITION. Un espace vectoriel muni d'une base de n vecteurs est dit *de dimension n* ou *n-dimensionnel.*

Dans l'espace nul il n'y a pas de base, car le vecteur nul forme un système lié. *La dimension de l'espace nul est* par définition *égale à zéro.*

Il peut arriver que, quel que soit l'entier naturel m, il existe dans l'espace $\mathscr{L}$ un système de m vecteurs linéairement indépendants. Cet espace est dit alors *de dimension infinie*. Il n'a pas de base.

EXEMPLES. 1) L'ensemble des vecteurs du plan représente un espace vectoriel de dimension deux, et l'ensemble de tous les vecteurs de l'espace, que l'on étudie en géométrie élémentaire, un espace vectoriel de dimension trois (voir propositions 7 et 8 du § 1, ch. I).

2) L'espace vectoriel des matrices-colonnes à n éléments est de dimension n. En effet, les colonnes de la matrice unité d'ordre n sont linéairement indépendantes (voir exemple, p. 126) et toute matrice-colonne à n éléments est leur combinaison linéaire (proposition 7, § 1, ch. V). L'espace vectoriel des matrices-colonnes à n éléments s'appelle *espace arithmétique de dimension n.*

3) L'espace vectoriel des fonctions d'une seule variable indépendante ξ, définies et continues pour $0 \leqslant \xi \leqslant 1$, est de dimension infinie. Pour le vérifier, il suffit de démontrer que pour tout m il existe dans cet espace m vecteurs linéairement indépendants. Soit un entier naturel m. Considérons m vecteurs (fonctions) de cet espace : $\xi^0 \equiv 1, \xi, \xi^2, \ldots, \xi^{m-1}$. Ils sont linéairement indépendants. En effet, si la combinaison linéaire de ces vecteurs avec coefficients $\alpha_0, \ldots, \alpha_{m-1}$ est égale au vecteur nul, le polynôme

$$\alpha_0 + \alpha_1\xi + \alpha_2\xi^2 + \ldots + \alpha_{m-1}\xi^{m-1}$$

est identiquement nul, ce qui n'est possible que si tous ses coefficients sont nuls.

L'algèbre linéaire étudie les espaces vectoriels de dimension finie. On supposera donc plus loin partout, à l'exception de quelques exemples, que l'espace est de dimension finie.

Dans un espace de dimension finie il existe une infinité de bases différentes. Ceci découle des propositions suivantes qu'on utilisera dans la suite.

PROPOSITION 8. *Dans un espace de dimension n, tout système ordonné de n vecteurs linéairement indépendants est une base.*

En effet, étant donné un tel système de vecteurs, chaque vecteur de l'espace se décompose suivant ces vecteurs, car, autrement, il y aurait dans l'espace $n + 1$ vecteurs linéairement indépendants.

PROPOSITION 9. *Dans un espace de dimension n tout système libre ordonné de $k < n$ vecteurs peut être complété jusqu'à une base.*

Cela découle du fait qu'à tout système de vecteurs linéairement indépendants on peut toujours ajouter un vecteur qui n'est pas leur combinaison linéaire. (S'il n'en était pas ainsi, le système serait lui-même une base.) On obtient donc $k + 1$ vecteurs linéairement indépendants. Si $k + 1 < n$, on reprend les raisonnements et on agit de la sorte jusqu'à ce qu'on n'obtienne n vecteurs linéairement indépendants, dont les k vecteurs donnés.

En particulier, on peut compléter jusqu'à la base tout vecteur non nul.

5. Changement de base. Etant donné deux bases $\|e_1, \ldots, e_n\|$ et $\|e_1', \ldots$ $\ldots, e_n'\|$ dans un espace de dimension n, on peut développer chaque vecteur de la base $\|e_1', \ldots, e_n'\|$ suivant les vecteurs $e_1, \ldots, e_n$:

$$e_i' = \sum_{j=1}^{n} \sigma_i^j e_j \quad (i = 1, \ldots, n). \tag{1}$$

Les composantes σ_i^j peuvent être écrites sous forme d'une matrice carrée d'ordre n :

$$S = \left\| \begin{matrix} \sigma_1^1 & \ldots & \sigma_n^1 \\ \sigma_1^2 & \ldots & \sigma_n^2 \\ \ldots & \ldots & \ldots \\ \sigma_1^n & \ldots & \sigma_n^n \end{matrix} \right\|.$$

Les colonnes de la matrice S représentent les colonnes de coordonnées des vecteurs $e_1', \ldots, e_n'$ dans la base e. Par suite, les colonnes de la matrice S sont linéairement indépendantes et $\det S \neq 0$.

DÉFINITION. La matrice dont la j-ième colonne représente la colonne de coordonnées du vecteur e_j' dans la base e s'appelle *matrice de passage* de la base e à la base e'.

L'égalité (1) peut être écrite en notations matricielles :

$$\|e'_1 \dots e'_n\| = \|e_1 \dots e_n\| S, \tag{2}$$

ou

$$e' = eS,$$

ce qui se vérifie aisément par multiplication de ces matrices.

En multipliant à droite les deux membres de l'égalité (2) par la matrice S^{-1}, on obtient

$$e = e'S^{-1},$$

d'où il découle que S^{-1} est la matrice de passage de e' à e.

PROPOSITION 10. *Etant donné la base e, toute matrice S de déterminant non nul est la matrice de passage de e à une base e'.*

En effet, si $\det S \neq 0$, les colonnes de la matrice S sont linéairement indépendantes. Ce sont les colonnes de coordonnées de n vecteurs linéairement indépendants constituant la base recherchée e'.

Voyons comment sont liées les composantes d'un même vecteur dans deux bases e et e'. Considérons le vecteur x et notons ξ et ξ' ses colonnes de coordonnées par rapport aux bases e et e'. Cela signifie en particulier que $x = e'\xi'$. Portons-y l'expression de e' par l'intermédiaire de e et de la matrice de passage de la base e à la base e'. On obtient $x = eS\xi'$. Or, $x = e\xi$. En comparant les deux dernières expressions, on a, en vertu de l'unicité du développement du vecteur x suivant les vecteurs de base,

$$\xi = S\xi'. \tag{3}$$

Sous une forme plus détaillée, cette formule peut être écrite

$$\left\|\begin{matrix}\xi^1 \\ \dots \\ \xi^n\end{matrix}\right\| = \left\|\begin{matrix}\sigma_1^1 \dots \sigma_n^1 \\ \dots\dots\dots \\ \sigma_1^n \dots \sigma_n^n\end{matrix}\right\| \left\|\begin{matrix}\xi'^1 \\ \dots \\ \xi'^n\end{matrix}\right\|$$

ou, si l'on effectue la multiplication des matrices,

$$\xi^i = \sum_{j=1}^{n} \sigma_j^i \xi'^j \quad (i = 1, \dots, n). \tag{4}$$

Ce résultat a déjà été obtenu pour l'espace de dimension trois (formule (2), § 4, ch. I).

§ 2. Sous-espace vectoriel

1. Définition et exemples. Dans un espace géométrique ordinaire, la somme des vecteurs d'un plan appartient aussi à ce plan, de même que le produit du vecteur par un nombre. Des propriétés analogues ont lieu pour

les vecteurs portés par une droite. Dans un espace vectoriel la généralisation du plan et de la droite est la notion de sous-espace vectoriel.

Définition. On dit qu'un ensemble non vide $\mathscr{L}'$ de vecteurs dans un espace vectoriel $\mathscr{L}$ est un *sous-espace vectoriel* si : a) la somme des vecteurs quelconques de $\mathscr{L}'$ appartient à $\mathscr{L}'$ et b) le produit de tout vecteur de $\mathscr{L}'$ par un nombre quelconque appartient aussi à $\mathscr{L}'$.

Le lecteur démontrera sans difficulté qu'en vertu de cette définition, toute combinaison linéaire de vecteurs de $\mathscr{L}'$ appartient à $\mathscr{L}'$. En particulier, le vecteur nul, en tant que produit $0x$, doit se trouver dans $\mathscr{L}'$. De même, pour tout vecteur x de $\mathscr{L}'$, le vecteur opposé, égal à $-1x$, se trouve dans $\mathscr{L}'$.

L'addition et la multiplication par un nombre, introduites dans l'espace $\mathscr{L}$, se définissent de la même façon dans son sous-espace $\mathscr{L}'$. Tous les axiomes de l'espace vectoriel sont vérifiés pour $\mathscr{L}'$ car ils le sont pour $\mathscr{L}$. Ainsi donc, tout sous-espace vectoriel a une structure d'espace vectoriel.

Exemple 1. Soit donné un ensemble $\mathscr{P}$ de vecteurs dans l'espace vectoriel $\mathscr{L}$. Notons $\mathscr{L}'$ l'ensemble de toutes les combinaisons linéaires dont chacune a un nombre fini de vecteurs appartenant à $\mathscr{P}$. L'ensemble $\mathscr{L}'$ est un sous-espace dans $\mathscr{L}$. En effet, si x et y appartiennent à $\mathscr{L}'$, on a

$$x = \sum_{i=1}^{n} \lambda_i p_i, \quad y = \sum_{j=1}^{n} \mu_j q_j,$$

où tous les p_i et q_j appartiennent à $\mathscr{P}$. On voit que $x + y = \sum \lambda_i p_i + \sum \mu_j q_j$, c'est-à-dire que la somme est toujours une combinaison linéaire d'un nombre fini de vecteurs de $\mathscr{P}$. D'une façon analogue, pour le vecteur $x = \sum \lambda_i p_i$ on a $\alpha x = \sum (\alpha\lambda_i) p_i$.

Le sous-espace $\mathscr{L}'$ ainsi construit s'appelle *enveloppe linéaire* de l'ensemble $\mathscr{P}$.

Soit $\{p_1, \ldots, p_m\}$ un système libre de vecteurs de $\mathscr{P}$ tel que chaque vecteur de $\mathscr{P}$ est une combinaison linéaire de ces vecteurs. (Si l'espace est de dimension finie, un tel système existe évidemment dans tout ensemble qui contient des vecteurs non nuls.) Les vecteurs $p_1, \ldots, p_m$ constituent une base dans l'enveloppe linéaire de l'ensemble $\mathscr{P}$. En effet, toute combinaison linéaire de vecteurs de $\mathscr{P}$ peut être représentée comme une combinaison linéaire des vecteurs $p_1, \ldots, p_m$, car il est toujours possible de développer chaque vecteur de $\mathscr{P}$ suivant $p_1, \ldots, p_m$ et de porter les développements obtenus dans la combinaison linéaire étudiée.

En particulier si $\mathscr{P}$ est un ensemble fini de vecteurs on a la

Proposition 1. *La dimension de l'enveloppe linéaire d'un ensemble fini de vecteurs ne dépasse pas le nombre de ces vecteurs.*

Exemple 2. Considérons un système homogène d'équations linéaires à n inconnues. L'ensemble de toutes les solutions de ce système représente un sous-espace de l'espace vectoriel des matrices-colonnes à n éléments.

Pour le démontrer, il suffit de se rappeler les propriétés des solutions du système homogène (proposition 3, § 5, ch. V).

Tout système fondamental de solutions du système d'équations considéré est une base de ce sous-espace.

Exemple 3. Dans tout espace vectoriel, l'ensemble constitué d'un seul vecteur nul est un sous-espace vectoriel. Ce sous-espace est dit *nul*.

Exemple 4. L'ensemble de tous les vecteurs de l'espace $\mathscr{L}$ est un sous-espace, c'est-à-dire que tout l'espace $\mathscr{L}$ est à la fois son sous-espace.

Nous laissons au lecteur le soin de vérifier les énoncés des deux derniers exemples.

Proposition 2. *Soit $\mathscr{L}'$ un sous-espace de l'espace vectoriel $\mathscr{L}_n$ de dimension n. La dimension de $\mathscr{L}'$ est alors $k \leqslant n$. Si $k = n$, $\mathscr{L}'$ coïncide avec $\mathscr{L}_n$.*

Admettons que $\mathscr{L}'$ est un sous-espace non nul, car c'est le seul cas qui exige une démonstration. Il existe alors dans $\mathscr{L}'$ un vecteur non nul à partir duquel on peut construire une base dans $\mathscr{L}'$, comme on l'a fait dans la démonstration de la proposition 9 du § 1. La construction doit s'achever au n-ième vecteur, vu que tout système libre dans $\mathscr{L}'$ est aussi libre dans $\mathscr{L}_n$ et par suite, ne peut contenir plus de n vecteurs.

Si la base dans $\mathscr{L}'$ contient n vecteurs, tout vecteur de $\mathscr{L}_n$ se développe suivant ces vecteurs et appartient donc à $\mathscr{L}'$. On voit donc que le sous-espace $\mathscr{L}'$ coïncide avec $\mathscr{L}_n$.

Etant donné que $\mathscr{L}'$ est l'enveloppe linéaire de sa base, il ressort de la proposition 2 que tout sous-espace d'un espace de dimension finie est une enveloppe linéaire d'un ensemble fini de vecteurs. Ceci se rapporte aussi au sous-espace nul car il est l'enveloppe linéaire du vecteur nul.

Proposition 3. *Soit $\mathscr{L}'$ un sous-espace de l'espace vectoriel n-dimensionnel $\mathscr{L}$ et soit $\|e_1, \ldots, e_k, e_{k+1}, \ldots, e_n\|$ une base dans $\mathscr{L}$, qui complète la base $\|e_1, \ldots, e_k\|$ dans $\mathscr{L}'$. Tous les vecteurs de $\mathscr{L}'$ sont alors les seuls vecteurs qui dans la base $\|e_1, \ldots, e_n\|$ possèdent les composantes $\xi^{k+1} = 0, \ldots, \xi^n = 0$.*

En effet, soit un vecteur x tel que $\xi^{k+1} = 0, \ldots, \xi^n = 0$. Il en ressort que $x = \xi^1 e_1 + \ldots + \xi^k e_k$ et, par suite, x appartient à $\mathscr{L}'$. Inversement, tout vecteur x de $\mathscr{L}'$ se développe en une combinaison linéaire $x = \xi^1 e_1 + \ldots + \xi^k e_k$ que l'on peut traiter comme un développement de x suivant les vecteurs $e_1, \ldots, e_n$ si l'on pose $\xi^{k+1} = 0, \ldots, \xi^n = 0$.

Remarquons que les égalités $\xi^{k+1} = 0, \ldots, \xi^n = 0$ peuvent être consi-

dérées comme un système d'équations linéaires reliant les coordonnées du vecteur x. Le rang de ce système est $n - k$ et ses solutions représentent les colonnes de coordonnées des vecteurs de $\mathscr{L}'$. On démontre aisément que $\mathscr{L}'$ se définit dans toute base par un système d'équations homogènes de rang $n - k$. En effet, dans tout changement de base, les anciennes composantes du vecteur s'expriment au moyen des nouvelles par les formules (4) du § 1, et dans la nouvelle base le système d'équations prend la forme

$$\sum_{i=1}^{n} \sigma_i^{k+1}\xi'^i = 0, \ldots, \sum_{i=1}^{n} \sigma_i^{n}\xi'^i = 0.$$

Le rang de ce système est toujours $n - k$, vu que les lignes de la matrice de passage sont linéairement indépendantes. Ainsi, on a démontré la

PROPOSITION 4. *Etant donné une base dans l'espace $\mathscr{L}$ de dimension n, les colonnes de coordonnées des vecteurs du sous-espace $\mathscr{L}'$ de dimension k vérifient le système d'équations linéaires homogènes de rang $n - k$**).

2. Somme et intersection de sous-espaces. Considérons deux sous-espaces $\mathscr{L}'$ et $\mathscr{L}''$ d'un espace vectoriel $\mathscr{L}$.

DÉFINITION. On appelle *somme* des sous-espaces $\mathscr{L}'$ et $\mathscr{L}''$, et on note $\mathscr{L}' + \mathscr{L}''$, l'enveloppe linéaire de leur réunion $\mathscr{L}' \cup \mathscr{L}''$.

D'une façon plus détaillée, la définition veut dire que tous les vecteurs x de $\mathscr{L}' + \mathscr{L}''$ (et eux seuls) peuvent être représentés sous forme de $x = \sum_i \alpha_i p_i + \sum_j \beta_j q_j$, où les vecteurs p_i appartiennent à $\mathscr{L}'$, et les vecteurs q_j à $\mathscr{L}''$. En désignant les sommes par x' et x'', on constate que le sous-espace $\mathscr{L}' + \mathscr{L}''$ se compose des vecteurs de la forme $x = x' + x''$, où x' est un vecteur de $\mathscr{L}'$ et x'' un vecteur de $\mathscr{L}''$.

Supposons que les dimensions des sous-espaces $\mathscr{L}'$ et $\mathscr{L}''$ sont k et l. Choisissons une base $\|e_1, \ldots, e_k\|$ dans $\mathscr{L}'$ et une base $\|f_1, \ldots, f_l\|$ dans $\mathscr{L}''$. Chaque vecteur de $\mathscr{L}' + \mathscr{L}''$ se décompose suivant les vecteurs $e_1, \ldots, e_k$, $f_1, \ldots, f_l$. On obtient une base dans $\mathscr{L}' + \mathscr{L}''$ en chassant de ce système de vecteurs tout vecteur qui s'exprime linéairement au moyen des autres. On le réalise ainsi.

Choisissons une base quelconque dans l'espace $\mathscr{L}$ et considérons une matrice formée des colonnes de coordonnées de tous les vecteurs $e_1, \ldots, e_k$, $f_1, \ldots, f_l$. Les vecteurs dont les colonnes de coordonnées contiennent le mineur principal de cette matrice constituent la base dans $\mathscr{L}' + \mathscr{L}''$. La démonstration n'étant pas difficile, on propose au lecteur de la faire à titre d'exercice.

---

*) Si $\mathscr{L}' = \mathscr{L}$, le système est de rang 0, autrement dit ne contient aucune équation non triviale.

DÉFINITION On appelle *intersection des sous-espaces* $\mathscr{L}'$ et $\mathscr{L}''$, et on note $\mathscr{L}' \cap \mathscr{L}''$, l'ensemble des vecteurs communs aux deux sous-espaces.

L'intersection $\mathscr{L}' \cap \mathscr{L}''$ *est un sous-espace.*

En effet, si les vecteurs x et y appartiennent à $\mathscr{L}' \cap \mathscr{L}''$, ils appartiennent à la fois à $\mathscr{L}'$ et à $\mathscr{L}''$. Il s'ensuit que les vecteurs $x + y$ et αx, avec un α quelconque, appartiennent de même à $\mathscr{L}'$ et à $\mathscr{L}''$, donc à $\mathscr{L}' \cap \mathscr{L}''$.

Si dans un espace $\mathscr{L}$ de dimension finie les sous-espaces $\mathscr{L}'$ et $\mathscr{L}''$ sont définis par des systèmes d'équations, leur intersection $\mathscr{L}' \cap \mathscr{L}''$ se définit par le système qui est la réunion de toutes les équations des systèmes qui définissent $\mathscr{L}'$ et $\mathscr{L}''$. Le système fondamental de solutions de ce système d'équations est alors une base dans le sous-espace $\mathscr{L}' \cap \mathscr{L}''$.

THÉORÈME 1. *La dimension de la somme de deux sous-espaces est égale à la somme de leurs dimensions moins la dimension de leur intersection.*

DÉMONSTRATION. Soient $\mathscr{L}'$ et $\mathscr{L}''$ deux sous-espaces dans l'espace $\mathscr{L}$ de dimension finie. Considérons dans $\mathscr{L}' + \mathscr{L}''$ le système de vecteurs défini de la façon suivante. Si l'intersection $\mathscr{L}' \cap \mathscr{L}''$ est un espace non nul, choisissons une base $\|e_1, \ldots, e_k\|$ dans $\mathscr{L}' \cap \mathscr{L}''$ et complétons-la par les vecteurs $f_1, \ldots, f_l$ jusqu'à la base de $\mathscr{L}'$ et par les vecteurs $g_1, \ldots, g_m$ jusqu'à la base de $\mathscr{L}''$. Si $\mathscr{L}' \cap \mathscr{L}''$ est un espace nul, prenons tout simplement la réunion d'une base de $\mathscr{L}'$ et d'une base de $\mathscr{L}''$. Démontrons préalablement que chaque vecteur x de $\mathscr{L}' + \mathscr{L}''$ est une combinaison linéaire des vecteurs choisis. En effet, $x = x' + x''$, où x' appartient à $\mathscr{L}'$, et x'' à $\mathscr{L}''$, si bien que x' se décompose suivant les vecteurs $f_1, \ldots, f_l$, $e_1, \ldots, e_k$, et x'' suivant les vecteurs $e_1, \ldots, e_k, g_1, \ldots, g_m$.

Montrons maintenant que ce système de vecteurs est libre. Soit une combinaison linéaire de ces vecteurs, égale à zéro :

$$\sum_{i=1}^{l} \alpha^i f_i + \sum_{j=1}^{k} \beta^j e_j + \sum_{s=1}^{m} \gamma^s g_s = o. \qquad (1)$$

Le vecteur $x = \sum \gamma^s g_s$ appartient à $\mathscr{L}''$. Or, il est évident que $x = - \sum \alpha^i f_i - \sum \beta^j e_j$ et, par suite, x appartient aussi à $\mathscr{L}'$. Ainsi, le vecteur x doit se trouver dans l'intersection $\mathscr{L}' \cap \mathscr{L}''$. D'où, en vertu de la proposition 3, on peut conclure que $\alpha^1 = \ldots = \alpha^l = 0$ et $\gamma^1 = \ldots = \gamma^m = 0$. Il ne reste dans l'égalité (1) que les termes

$$\beta^1 e_1 + \ldots + \beta^k e_k$$

si $\mathscr{L}' \cap \mathscr{L}'' \neq o$. Mais leurs coefficients sont aussi nuls car les vecteurs $e_1, \ldots, e_k$ sont linéairement indépendants. Ainsi donc, la combinaison linéaire (1) est nécessairement triviale et tous les vecteurs sont linéairement

indépendants. On a montré que le système des vecteurs $f_1, \ldots, f_l, e_1, \ldots, e_k, g_1, \ldots, g_m$ est une base dans le sous-espace $\mathscr{L}' + \mathscr{L}''$.

On peut maintenant achever sans difficulté la démonstration. La dimension de $\mathscr{L}'$ est $l + k$, celle de $\mathscr{L}''$, $k + m$, tandis que celle de la somme $\mathscr{L}' + \mathscr{L}''$ est, comme on vient de montrer, $k + l + m$.

Il découle en particulier du théorème 1 que deux sous-espaces d'un espace de dimension n ont toujours une intersection non nulle si la somme des dimensions de ces sous-espaces est strictement supérieure à n. En effet, la dimension de leur somme ne peut dépasser n.

3. Somme directe de sous-espaces. Si l'intersection des sous-espaces $\mathscr{L}'$ et $\mathscr{L}''$ est le sous-espace nul, leur somme $\mathscr{L}' + \mathscr{L}''$ est appelée *somme directe* et est notée $\mathscr{L}' \dotplus \mathscr{L}''$ ou $\mathscr{L}' \oplus \mathscr{L}''$.

La dimension de la somme directe est égale à la somme des dimensions des termes. Soient $\|f_1, \ldots, f_l\|$ une base dans $\mathscr{L}'$ et $\|g_1, \ldots, g_m\|$ une base dans $\mathscr{L}''$. De la démonstration du théorème 1 il découle que le système des vecteurs $f_1, \ldots, f_l, g_1, \ldots, g_m$ est une base dans $\mathscr{L}' \oplus \mathscr{L}''$.

PROPOSITION 5. *Tout vecteur x de la somme directe $\mathscr{L}' \oplus \mathscr{L}''$ se décompose de façon unique en somme des vecteurs $x' \in \mathscr{L}'$ et $x'' \in \mathscr{L}''$.*

En effet, s'il y avait deux décompositions différentes $x = x' + x''$ et $x = y' + y''$, on aurait $x' - y' = y'' - x''$. Il est clair que le vecteur $x' - y' \in \mathscr{L}'$. Or il est égal à $y'' - x''$ et, par suite, appartient à $\mathscr{L}''$. Donc, $x' - y' \in \mathscr{L}' \cap \mathscr{L}''$, d'où $x' = y'$ et $x'' = y''$.

Signalons en conclusion de ce paragraphe que les notions de somme et d'intersection de deux sous-espaces peuvent être facilement étendues à toute famille finie de sous-espaces.

DÉFINITION. On dit que la somme des sous-espaces $\mathscr{L}^{(1)}, \ldots, \mathscr{L}^{(s)}$ est *directe* et on la note $\mathscr{L}^{(1)} \oplus \ldots \oplus \mathscr{L}^{(s)}$ ou $\bigoplus_{i=1}^{s} \mathscr{L}^{(i)}$ si

$$\mathscr{L}^{(k)} \cap \sum_{i \neq k} \mathscr{L}^{(i)} = o \text{ pour tous les } k = 1, \ldots, s.$$

On voit immédiatement que dans ce cas

$$\mathscr{L}^{(1)} \oplus \ldots \oplus \mathscr{L}^{(j)} = (\mathscr{L}^{(1)} \oplus \ldots \oplus \mathscr{L}^{(j-1)}) + \mathscr{L}^{(j)},$$
$$j = 2, \ldots, s.$$

En se servant de cette formule et des propriétés de la somme directe de deux sous-espaces, on démontre aisément par récurrence que la réunion des bases des sous-espaces $\mathscr{L}^{(i)}$ est une base de leur somme directe. Il s'ensuit en particulier que la dimesion de la somme directe est égale à la somme des dimensions des termes.

Chaque vecteur de la somme des sous-espaces se décompose en une somme $\sum x_i$, où $x_i \in \mathscr{L}^{(i)}$. Si la somme est directe, cette décomposition est unique. En effet, si un vecteur x admettait deux décompositions différentes $x = \sum x_i = \sum y_i$, on aurait $\sum (x_i - y_i) = o$. Or cette égalité entraîne immédiatement que les vecteurs de la réunion des bases des sous-espaces $\mathscr{L}^{(i)}$ sont linéairement dépendants.

Laissons au lecteur le soin de vérifier que toutes les propriétés de la somme directe mentionnées ici sont équivalentes à la définition de la somme directe.

§ 3. Applications linéaires

1. Définition. Soient $\mathscr{L}$ et $\widetilde{\mathscr{L}}$ deux espaces vectoriels, tous deux réels ou complexes. On appelle *application* **A** de l'espace $\mathscr{L}$ dans l'espace $\widetilde{\mathscr{L}}$ une relation qui associe à chaque vecteur de $\mathscr{L}$ un vecteur de $\widetilde{\mathscr{L}}$ et un seul. On le note tout court $\mathbf{A} : \mathscr{L} \to \widetilde{\mathscr{L}}$. L'image du vecteur x est notée $\mathbf{A}(x)$.

DÉFINITION. L'application $\mathbf{A} : \mathscr{L} \to \widetilde{\mathscr{L}}$ est dite *linéaire* si pour tous vecteurs x et y de $\mathscr{L}$ et tout nombre α sont vérifiées les égalités

$$\mathbf{A}(x + y) = \mathbf{A}(x) + \mathbf{A}(y), \quad \mathbf{A}(\alpha x) = \alpha \mathbf{A}(x). \tag{1}$$

Il faut souligner que le signe « + » dans les deux membres de la première formule de (1) désigne en général deux opérations différentes : l'addition dans l'espace $\mathscr{L}$ et l'addition dans l'espace $\widetilde{\mathscr{L}}$. Une remarque analogue peut être faite à propos de la seconde formule.

Il découle immédiatement de la définition que l'image par application linéaire d'une combinaison linéaire de vecteurs est la combinaison linéaire de leurs images.

On dit que l'application linéaire est une *transformation* linéaire si les espaces $\mathscr{L}$ et $\widetilde{\mathscr{L}}$ se confondent.

Voici quelques exemples d'applications linéaires.

1) Soit λ un nombre fixé. Associons à chaque vecteur x de l'espace $\mathscr{L}$ le vecteur λx. Il est aisé de voir que c'est une transformation linéaire.

2) Dans une transformation affine du plan, l'espace bidimensionnel des vecteurs de ce plan s'applique sur lui-même. Ceci étant, l'image de la somme de vecteurs est la somme de leurs images, tandis que l'image du produit d'un vecteur par un nombre est le produit de l'image du vecteur par ce nombre.

3) Choisissons une base dans un espace vectoriel réel $\mathscr{L}_n$ de dimension n et associons à chaque vecteur sa colonne de coordonnées. L'application ainsi définie est une application linéaire de l'espace considéré dans l'espace des matrices-colonnes à n éléments.

Si l'on fait correspondre à chaque vecteur sa première composante, on obtient une application de l'espace $\mathscr{L}_n$ dans l'espace vectoriel $\mathscr{R}$ des nom-

bres réels. Cette application est linéaire en vertu des propriétés des opérations sur les composantes des vecteurs. Les applications linéaires de $\mathscr{L}_n$ dans $\mathscr{R}$ s'appellent *fonctions linéaires* sur $\mathscr{L}_n$. On les étudiera plus loin dans le chapitre VIII.

4) Soient $C^0[-1, 1]$ et $C^0[0, 2]$ les espaces de fonctions continues respectivement sur les segments $[-1, 1]$ et $[0, 2]$. Faisons correspondre à toute fonction $f(x)$ de $C^0[-1, 1]$ la fonction $f(x + 1)$ de $C^0[0, 2]$. L'application ainsi définie est évidemment linéaire. Un exemple moins trivial peut être obtenu en faisant correspondre à chaque fonction de $C^0[-1, 1]$ sa fonction primitive $F(x)$ satisfaisant à la condition $F(0) = 0$.

5) Considérons l'espace arithmétique $\mathscr{R}_n$ de dimension n (espace des matrices-colonnes à n éléments) et une matrice A à m lignes et n colonnes. Associons à chaque matrice-colonne ξ de $\mathscr{R}_n$ une matrice-colonne $A\xi$ qui comprend m éléments. La relation ainsi définie est une application de $\mathscr{R}_n$ dans $\mathscr{R}_m$. Cette application est linéaire en vertu des propriétés de la multiplication des matrices.

6) L'application qui fait correspondre à chaque vecteur le vecteur nul est linéaire. On l'appelle *application nulle*.

Dans la suite de ce paragraphe, les lettres n et m désigneront les dimensions respectives des espaces $\mathscr{L}$ et $\tilde{\mathscr{L}}$.

Démontrons la propriété générale suivante des applications linéaires.

PROPOSITION 1. *Etant donné une application linéaire* $\mathbf{A} : \mathscr{L}_n \to \tilde{\mathscr{L}}_m$, *l'image de tout sous-espace vectoriel* $\mathscr{L}'$ *de* $\mathscr{L}_n$ *est un sous-espace vectoriel* $\mathbf{A}(\mathscr{L}')$ *de* $\tilde{\mathscr{L}}_m$ *dont la dimension ne dépasse pas celle de* $\mathscr{L}'$.

En effet, soit $\|e_1, \ldots, e_k\|$ une base dans $\mathscr{L}'$. Pour tout vecteur x de $\mathscr{L}'$ on a $x = \xi^1 e_1 + \ldots + \xi^k e_k$ et, par suite,

$$\mathbf{A}(x) = \mathbf{A}(\xi^1 e_1 + \ldots + \xi^k e_k) = \xi^1 \mathbf{A}(e_1) + \ldots + \xi^k \mathbf{A}(e_k). \tag{2}$$

Cela signifie que tout élément de l'ensemble $\mathbf{A}(\mathscr{L}')$ des images de tous les vecteurs de $\mathscr{L}'$ est une combinaison linéaire des vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_k)$. Inversement, toute combinaison linéaire des vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_k)$ est évidemment l'image d'un vecteur de $\mathscr{L}'$. Ainsi, l'ensemble $\mathbf{A}(\mathscr{L}')$ se confond avec l'enveloppe linéaire des vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_k)$ et partant, est un sous-espace. La dimension de ce sous-espace ne dépasse pas k en vertu de la proposition 1 du § 2.

Il faut signaler un cas particulier de la proposition démontrée : l'ensemble des images de tous les vecteurs de $\mathscr{L}_n$ est un sous-espace dans $\tilde{\mathscr{L}}_m$.

On désignera ce sous-espace par $\mathbf{A}(\mathscr{L}_n)$ en l'appelant *ensemble des valeurs* de l'application.

DÉFINITION. La dimension de l'ensemble des valeurs de l'application $\mathbf{A}$ est appelée *rang* de cette application.

Si le rang de $\mathbf{A}$ est égal à m, c'est-à-dire si $\mathbf{A}(\mathscr{L}_n)$ coïncide avec $\widetilde{\mathscr{L}}_m$, chaque vecteur de $\widetilde{\mathscr{L}}_m$ est l'image d'un vecteur de $\mathscr{L}_n$. On dit alors que $\mathbf{A}$ est une *application surjective*, ou *surjection*.

PROPOSITION 2. *L'ensemble des vecteurs de $\mathscr{L}_n$ dont l'image par l'application $\mathbf{A}$ est le vecteur nul est un sous-espace vectoriel dans $\mathscr{L}_n$.*

En effet, si deux vecteurs se transforment en vecteur nul, il en est de même de leur somme. Si $\mathbf{A}(x) = o$, on a $(\alpha x) = \alpha o = o$.

DÉFINITION. On appelle *noyau* de l'application $\mathbf{A}$ un sous-espace de vecteurs dont l'image est le vecteur nul.

Le noyau d'une application n'est jamais un ensemble vide car il contient au moins le vecteur nul. En effet, $\mathbf{A}(o) = \mathbf{A}(0x) = 0\mathbf{A}(x) = o$. Si la dimension du noyau est différente de zéro et le noyau contient au moins un vecteur non nul, il existe dans $\widetilde{\mathscr{L}}_m$ des vecteurs qui admettent au moins deux antécédents (tel est le cas du vecteur nul de $\widetilde{\mathscr{L}}_m$). La réciproque est également vraie : s'il existe un vecteur $\tilde{x}$ ayant deux antécédents différents, c'est-à-dire si $\mathbf{A}(x) = \mathbf{A}(y) = \tilde{x}$, le noyau de $\mathbf{A}$ contient un vecteur non nul. En effet, on a dans ce cas $x - y \neq 0$ et $\mathbf{A}(x - y) = o$.

L'application dans laquelle les images des vecteurs distincts sont distinctes est appelée *application injective* ou *injection*. On a donc la

PROPOSITION 3. *Une application est injective si et seulement si son noyau est le sous-espace nul.*

2. Expression analytique d'une application linéaire. Considérons deux espaces vectoriels $\mathscr{L}_n$ et $\widetilde{\mathscr{L}}_m$ de dimensions n et m et une application $\mathbf{A} : \mathscr{L}_n \to \widetilde{\mathscr{L}}_m$. Soit $\|e_1, \ldots, e_n\|$ une base dans l'espace $\mathscr{L}_n$. On peut alors représenter l'image d'un vecteur quelconque $x = \xi^1 e_1 + \ldots + \xi^n e_n$ sous la forme

$$\mathbf{A}(x) = \xi^1 \mathbf{A}(e_1) + \ldots + \xi^n \mathbf{A}(e_n). \tag{3}$$

Donc, $\mathbf{A}(x)$ peut être défini d'après les composantes de x si sont connus les n vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_n)$ dans $\widetilde{\mathscr{L}}_m$.

Soit $\|f_1, \ldots, f_m\|$ une base dans l'espace $\widetilde{\mathscr{L}}_m$. Décomposons chacun des vecteurs $\mathbf{A}(e_i)$ suivant les vecteurs de cette base :

$$\mathbf{A}(e_i) = \sum_{p=1}^{m} \alpha_i^p f_p \quad (i = 1, \ldots, n).$$

Si les composantes de $\mathbf{A}(x)$ par rapport à la base f sont notées $\eta^1, \ldots$ $\ldots, \eta^m$, l'égalité (3) peut être écrite ainsi :

$$\sum_{p=1}^{m} \eta^p f_p = \sum_{i,p} \xi^i \alpha_i^p f_p.$$

D'où, en vertu de l'unicité de la décomposition dans une base :

$$\eta^p = \sum_{i=1}^{n} \alpha_i^p \xi^i \tag{4}$$

pour tout $p = 1, \ldots, m$.

Soit A une matrice à éléments α_i^p. L'égalité (4) peut alors être écrite sous la forme matricielle :

$$\eta = A\xi \tag{5}$$

ou de façon plus détaillée

$$\begin{Vmatrix} \eta^1 \\ \ldots \\ \eta^m \end{Vmatrix} = \begin{Vmatrix} \alpha_1^1 & \ldots & \alpha_n^1 \\ \ldots & \ldots & \ldots \\ \alpha_1^m & \ldots & \alpha_n^m \end{Vmatrix} \begin{Vmatrix} \xi^1 \\ \ldots \\ \xi^n \end{Vmatrix}.$$

La colonne de coordonnées de l'image (dans la base f) est exprimée ici sous forme de produit de la matrice (m, n) par la colonne de coordonnées de l'antécédent (dans la base e). Il est utile de comparer ce résultat avec l'exemple 5) du point précédent.

DÉFINITION. On appelle *matrice de l'application linéaire* $\mathbf{A} : \mathscr{L}_n \to \tilde{\mathscr{L}}_m$ par rapport aux bases e et f, la matrice dont les colonnes (prises dans leur ordre naturel) représentent les colonnes de coordonnées des vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_n)$ dans la base f.

La matrice A, construite plus haut avec les éléments α_i^p, est justement la matrice de l'application $\mathbf{A}$ par rapport aux bases choisies.

La matrice de l'application linéaire est parfaitement définie au sens suivant : si pour tout vecteur $x = e\xi$ la colonne de coordonnées de l'image $\mathbf{A}(x)$ dans la base f est $\eta = B\xi$, les colonnes de la matrice B sont des colonnes de coordonnées des vecteurs $\mathbf{A}(e_j)$, de sorte que la matrice B coïncide avec la matrice A.

Cette assertion est facile à vérifier. Multiplions la matrice B par la colonne de coordonnées du vecteur e_i, c'est-à-dire par la i-ième colonne de la matrice unité d'ordre n. Il est évident que le produit est égal à la i-ième colonne de B qui est la colonne de coordonnées de $\mathbf{A}(e_i)$.

L'exemple 5) montre que, les bases e et f étant choisies, toute matrice de type (m, n) est la matrice d'une application linéaire $\mathscr{L}_n \to \tilde{\mathscr{L}}_m$.

Ainsi donc, on voit que le choix de bases dans les espaces $\mathscr{L}_n$ et $\tilde{\mathscr{L}}_m$ établit une correspondance biunivoque entre les applications $\mathscr{L}_n \to \tilde{\mathscr{L}}_m$ et les matrices de type (m, n).

PROPOSITION 4. *Le rang de la matrice d'une application linéaire est égal au rang de cette application.*

DÉMONSTRATION. Soient $j_1, \ldots, j_r$ les numéros des colonnes renfermant le mineur principal de la matrice A d'une application linéaire **A**. Cela signifie que les vecteurs $\mathbf{A}(e_{j_1}), \ldots, \mathbf{A}(e_{j_r})$ sont linéairement indépendants et que chaque vecteur $\mathbf{A}(e_i)$ $(i = 1, \ldots, n)$ est leur combinaison linéaire. Par suite, on est en mesure d'exprimer l'image de tout vecteur $\mathbf{A}(x)$ au moyen des seuls vecteurs $\mathbf{A}(e_{j_1}), \ldots, \mathbf{A}(e_{j_r})$. Donc, ces vecteurs constituent une base dans l'ensemble des valeurs de l'application **A**, et leur nombre est égal à la dimension de $\mathbf{A}(\mathscr{L}_n)$, c'est-à-dire au rang de l'application. La proposition est démontrée.

Il résulte de la proposition 4 que le rang de la matrice d'une application linéaire est le même, quel que soit le couple de bases choisi.

Profitons de l'expression analytique d'une application linéaire pour démontrer la proposition suivante.

PROPOSITION 5. *La somme du rang de l'application* $\mathbf{A} : \mathscr{L}_n \to \tilde{\mathscr{L}}_m$ *et de la dimension de son noyau est égale à la dimension de l'espace* $\mathscr{L}_n$.

DÉMONSTRATION. Selon la formule (5), le noyau de l'application se définit par le système d'équations linéaires homogènes $A\xi = o$ à n inconnues. Comme il découle de la proposition 4, le rang de la matrice de ce système est égal au rang r de l'application. Soit d la dimension du noyau. La propriété de l'ensemble des solutions du système d'équations homogènes entraîne alors que $d = n - r$. La proposition est démontrée.

En particulier, l'égalité $r = n$ est nécessaire et suffisante pour que l'application ait un noyau nul, c'est-à-dire pour qu'elle soit une injection.

Rappelons que l'application est dite *bijective* si tout vecteur $\tilde{x}$ de $\tilde{\mathscr{L}}_m$ est l'image d'un vecteur x de $\mathscr{L}_n$ et d'un seul. Cela signifie qu'une application bijective est en même temps une injection et une surjection. Pour une injection on a $r = n$ et pour une surjection, $r = m$. Ainsi donc, on a la

PROPOSITION 6. *L'application* $\mathbf{A} : \mathscr{L}_n \to \tilde{\mathscr{L}}_m$ *est bijective si et seulement si les dimensions des espaces coïncident et sont égales au rang de l'application.*

Cette proposition résulte aussi facilement à partir de l'étude du système d'équations linéaires (4) : le fait que l'application est bijective signifie que ce système possède une solution unique ξ pour toute matrice-colonne des termes constants η.

3. Isomorphisme d'espaces linéaires. DÉFINITION. On appelle *isomorphisme* une application linéaire bijective d'un espace vectoriel sur l'autre. S'il existe un isomorphisme de $\mathscr{L}$ sur $\tilde{\mathscr{L}}$, les espaces $\mathscr{L}$ et $\tilde{\mathscr{L}}$ sont dits *isomorphes*.

Il découle de la proposition 6 que, pour que deux espaces soient isomorphes, il faut que leurs dimensions coïncident. Il s'avère que cette condition

est également suffisante, c'est-à-dire qu'a lieu le théorème d'isomorphisme suivant.

THÉORÈME 1. Deux espaces vectoriels réels (*resp. complexes*) *sont isomorphes si et seulement si leurs dimensions sont égales.*

Il nous reste à démontrer que la condition est suffisante. Soient $\mathscr{L}$ et $\widetilde{\mathscr{L}}$ deux espaces vectoriels de dimension n. Si une base est choisie dans chacun d'eux, toute matrice carrée d'ordre n définit une application de $\mathscr{L}$ dans $\widetilde{\mathscr{L}}$ par la formule (5). Cette application est un isomorphisme si le rang de la matrice est égal à n. Ainsi, pour définir un isomorphisme de $\mathscr{L}$ sur $\widetilde{\mathscr{L}}$, il suffit de choisir des bases dans ces espaces et de définir une matrice de rang n (c'est-à-dire de déterminant non nul).

L'importance du théorème d'isomorphisme est la suivante. Lorsqu'on étudie les propriétés des espaces vectoriels qui sont liées aux opérations d'addition et de multiplication par un nombre, on ne s'intéresse pas à la nature de leurs éléments, que ce soient colonnes, polynômes, nombres, segments orientés, fonctions ou matrices. Les propriétés de deux espaces isomorphes sont absolument identiques. Au point de vue algébrique, les espaces isomorphes sont identiques. Si l'on convient de ne pas différencier les espaces isomorphes, le théorème d'isomorphisme nous permet d'associer à chaque dimension un seul espace vectoriel.

4. Variation de la matrice d'une application linéaire avec le changement de base. Considérons une application linéaire $\mathbf{A} : \mathscr{L}_n \to \widetilde{\mathscr{L}}_m$. Les espaces $\mathscr{L}_n$ et $\widetilde{\mathscr{L}}_m$ étant rapportés aux bases respectives $\boldsymbol{e}$ et $\boldsymbol{f}$, $\mathbf{A}$ se définit par la matrice A. Soit un autre couple de bases $\boldsymbol{e}'$ et $\boldsymbol{f}'$ liées à $\boldsymbol{e}$ et $\boldsymbol{f}$ par des matrices de passage S et P, et soit A' la matrice de l'application $\mathbf{A}$ par rapport aux bases $\boldsymbol{e}'$ et $\boldsymbol{f}'$. Il nous faut trouver une relation entre les matrices A et A'.

Considérons un vecteur x de l'espace $\mathscr{L}_n$ et son image $y = \mathbf{A}(x)$. Désignons les colonnes de coordonnées de x dans les bases $\boldsymbol{e}$ et $\boldsymbol{e}'$ respectivement par ξ et ξ', et les colonnes de coordonnées du vecteur y dans les bases $\boldsymbol{f}$ et $\boldsymbol{f}'$ par η et η'. Selon la formule (3) du § 1, on a les relations matricielles suivantes entre les anciennes et les nouvelles coordonnées des vecteurs x et y :

$$\xi = S\xi', \qquad \eta = P\eta'.$$

En portant ces expressions dans la formule (5), on obtient $P\eta' = AS\xi'$. Vu que la matrice de passage possède toujours son inverse, on peut définir η' à partir de cette égalité : $\eta' = P^{-1}AS\xi'$. Or, $\eta' = A'\xi'$ par définition de A'. La matrice de l'application linéaire étant parfaitement définie par rapport aux bases données, il vient

$$A' = P^{-1}AS. \tag{6}$$

C'est la relation cherchée entre les matrices A et A'.

Si l'on désigne les éléments des matrices A et A' par α^i_j et α'^i_j, et les éléments de P^{-1} et S respectivement par ρ^i_k et σ^l_j, l'égalité matricielle (6) peut être écrite sous forme de mn égalités numériques :

$$\alpha'^i_j = \sum_{k,l} \rho^i_k \alpha^k_l \sigma^l_j \tag{7}$$

Les indices prennent ici les valeurs suivantes : $i, k = 1, \ldots, m$; $j, l = 1, \ldots$ $\ldots, n$.

5. Forme canonique d'une matrice de l'application linéaire. Il découle de la formule (6) que la matrice de l'application linéaire $\mathbf{A} : \mathscr{L}_n \to \widetilde{\mathscr{L}}_m$ varie avec le changement de bases dans $\mathscr{L}_n$ et $\widetilde{\mathscr{L}}_m$. Il se pose tout naturellement la question : comment choisir les bases dans $\mathscr{L}_n$ et $\widetilde{\mathscr{L}}_m$ pour que la matrice de l'application linéaire ait la plus simple forme.

THÉORÈME 2. *Pour toute application linéaire $\mathbf{A} : \mathscr{L}_n \to \widetilde{\mathscr{L}}_m$ on peut choisir des bases e et f dans les espaces $\mathscr{L}_n$ et $\widetilde{\mathscr{L}}_m$ de manière que la matrice de l'application soit de la forme* *)

$$A = \left\| \begin{array}{c|c} E_r & O \\ \hline O & O \end{array} \right\|, \tag{8}$$

où E_r est la matrice unité d'ordre r.

DÉMONSTRATION. Soit r le rang de l'application $\mathbf{A}$. Choisissons une base e dans l'espace $\mathscr{L}_n$ de la façon suivante : soient $e_{r+1}, \ldots, e_n$ ses $n - r$ vecteurs qui appartiennent au noyau de l'application $\mathbf{A}$ (sa dimension est justement $n - r$), et soient $e_1, \ldots, e_r$ des vecteurs choisis arbitrairement. En vertu de ce choix, les $n - r$ dernières colonnes de la matrice sont nulles quelle que soit la base f dans l'espace $\widetilde{\mathscr{L}}_m$. Vu que le rang de la matrice est r, les r premières colonnes doivent être linéairement indépendantes. Cela signifie que les vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_r)$ sont linéairement indépendants. Admettons qu'ils soient les r premiers vecteurs de base dans l'espace $\widetilde{\mathscr{L}}_m$: $f_1 = \mathbf{A}(e_1), \ldots, f_r = \mathbf{A}(e_r)$, les autres vecteurs $f_{r+1}, \ldots, f_m$ pouvant être choisis arbitrairement. Avec ce choix de la base, les r premières colonnes de la matrice sont les r premières colonnes de la matrice unité d'ordre m. C'est justement la forme (8) de la matrice de l'application.

6. Somme et produit de deux applications. Considérons deux applications linéaires $\mathbf{A} : \mathscr{L} \to \widetilde{\mathscr{L}}$ et $\mathbf{B} : \mathscr{L} \to \widetilde{\mathscr{L}}$. On appelle *somme* des applications $\mathbf{A}$ et $\mathbf{B}$, et on note $\mathbf{A} + \mathbf{B}$, l'application $\mathbf{C} : \mathscr{L} \to \widetilde{\mathscr{L}}$ définie par l'égalité $\mathbf{C}(x) = \mathbf{A}(x) + \mathbf{B}(x)$. On vérifie aisément que $\mathbf{C}$ est une application linéaire. En effet, si des bases sont choisies dans $\mathscr{L}$ et $\widetilde{\mathscr{L}}$, les colonnes de coordonnées des vecteurs $\mathbf{A}(x)$ et $\mathbf{B}(x)$ s'expriment au moyen des matrices

---

*) Si $r = m$ (resp. $r = n$), la matrice (8) n'a pas de lignes (resp. colonnes) nulles.

des applications sous la forme $A\xi$ et $B\xi$. Par suite, la colonne de coordonnées de $\mathbf{C}(x)$ est de la forme $A\xi + B\xi = (A + B)\xi$. Donc, $\mathbf{A} + \mathbf{B}$ est une application linéaire et sa matrice est égale à la somme des matrices des applications $\mathbf{A}$ et $\mathbf{B}$.

Le produit de l'application linéaire $\mathbf{A} : \mathscr{L} \to \tilde{\mathscr{L}}$ par le nombre α se définit comme application $\mathbf{B} : \mathscr{L} \to \tilde{\mathscr{L}}$ qui fait correspondre à tout vecteur x le vecteur $\alpha\mathbf{A}(x)$. Il n'est pas difficile de vérifier qu'elle est linéaire et possède la matrice αA si A est la matrice de $\mathbf{A}$.

Le résultat de deux applications successives $\mathbf{A} : \mathscr{L} \to \tilde{\mathscr{L}}$ et $\mathbf{B} : \tilde{\mathscr{L}} \to \tilde{\tilde{\mathscr{L}}}$ est appelé leur *produit* et est noté $\mathbf{B} \circ \mathbf{A}$ (la première application est écrite à droite). Il va de soi que $\mathbf{B} \circ \mathbf{A}$ est une application linéaire de $\mathscr{L}$ dans $\tilde{\tilde{\mathscr{L}}}$.

Soient $\mathscr{L}$, $\tilde{\mathscr{L}}$ et $\tilde{\tilde{\mathscr{L}}}$ des espaces vectoriels, et $\boldsymbol{e}$, $\boldsymbol{f}$ et $\boldsymbol{g}$ leurs bases respectives. Désignons par A la matrice de l'application $\mathbf{A}$ par rapport aux bases $\boldsymbol{e}$ et $\boldsymbol{f}$ et par B la matrice de $\mathbf{B}$ par rapport aux bases $\boldsymbol{f}$ et $\boldsymbol{g}$.

PROPOSITION 7. *L'application* $\mathbf{B} \circ \mathbf{A}$ *admet* BA *pour matrice par rapport aux bases* $\boldsymbol{e}$ *et* $\boldsymbol{g}$.

Pour le démontrer, considérons la colonne des coordonnées ξ du vecteur arbitraire x de $\mathscr{L}$. Notons η et ζ les colonnes de coordonnées respectives des vecteurs $\mathbf{A}(x)$ et $\mathbf{B}(\mathbf{A}(x))$. Selon la formule (5), on a alors $\eta = A\xi$ et $\zeta = B\eta = BA\xi$, ce qu'il fallait démontrer.

Puisque le rang de l'application coïncide avec celui de sa matrice, on a, conformément à la proposition 5 du § 6, ch. V sur le rang du produit des matrices, la proposition suivante.

PROPOSITION 8. *Le rang du produit des applications ne dépasse pas ceux des facteurs.*

Le produit de l'application $\mathbf{A}$ par un nombre peut être assimilé au produit de $\mathbf{A}$ par une transformation linéaire qui consiste en multiplication de tous les vecteurs par ce nombre (voir exemple 1), p. 182).

Les propriétés de l'addition et de la multiplication des applications découlent immédiatement des propriétés correspondantes de la multiplication des matrices et on ne s'y arrêtera pas. Laissons au lecteur le soin de formuler et de démontrer, par exemple, la propriété d'associativité de la multiplication des applications.

Soit donnée une application linéaire $\mathbf{A} : \mathscr{L}_n \to \tilde{\mathscr{L}}_m$. L'application linéaire $\mathbf{B} : \tilde{\mathscr{L}}_m \to \mathscr{L}_n$ est dite *réciproque* de $\mathbf{A}$ et notée $\mathbf{A}^{-1}$ si $\mathbf{B} \circ \mathbf{A} = \mathbf{E}$ et $\mathbf{A} \circ \mathbf{B} = \tilde{\mathbf{E}}$, où $\mathbf{E}$ et $\tilde{\mathbf{E}}$ sont des transformations identiques des espaces $\mathscr{L}_n$ et $\tilde{\mathscr{L}}_m$. Autrement dit, pour chaque x de $\mathscr{L}_n$ on doit avoir $\mathbf{B}(\mathbf{A}(x)) = x$, et $\mathbf{A}(\mathbf{B}(y)) = y$ pour tout y de $\tilde{\mathscr{L}}_m$.

PROPOSITION 9. *L'application linéaire* $\mathbf{A}$ *admet une réciproque si et seulement si* $\mathbf{A}$ *est un isomorphisme.*

DÉMONSTRATION. Soit **A** un isomorphisme. Son rang satisfait alors à la condition $r = m = n$ et, par suite, on peut appliquer la règle de Cramer au système d'équations $A\xi = \eta$ liant les coordonnées de l'image à son antécédent dans un couple de bases. Ce système a une solution unique pour toute matrice-colonne de termes constants. Considérons une application **B** : $\tilde{\mathscr{L}}_m \to \mathscr{L}_n$ qui à tout vecteur y avec colonne de coordonnées η associe le vecteur x avec colonne de coordonnées ξ égale à la solution $A^{-1}\eta$ de ce système. Vu que la matrice-colonne $\xi = A^{-1}\eta$ est égale au produit de η par une matrice, l'application **B** est linéaire. Il va de soi qu'on a toujours $\mathbf{B}(\mathbf{A}(x)) = x$ et $\mathbf{A}(\mathbf{B}(y)) = y$. Donc, l'application **B** est la réciproque de **A**.

Supposons maintenant que **A** n'est pas un isomorphisme. On a alors soit $r < m$, soit $r < n$. Dans le premier cas, il existe dans $\tilde{\mathscr{L}}_m$ un vecteur u qui n'appartient pas à $\mathbf{A}(\mathscr{L}_n)$. Si l'application réciproque existait, on aurait $u = \mathbf{A}(\mathbf{A}^{-1}(u)) \in \mathbf{A}(\mathscr{L}_n)$, ce qui contredit le choix de u. Dans le second cas, il existe un vecteur $z \neq o$ appartenant au noyau de **A**. Si l'application réciproque existait, on obtiendrait l'égalité $z = \mathbf{A}^{-1}(\mathbf{A}(z)) = o$ contredisant le choix de z. La proposition est démontrée.

En démontrant la première partie de la proposition on a obtenu la matrice de l'application $\mathbf{A}^{-1}$. Plus précisément, si la matrice de l'application **A** par rapport aux bases $\boldsymbol{e}$ et $\boldsymbol{f}$ est A, celle de l'application $\mathbf{A}^{-1}$ par rapport aux bases $\boldsymbol{f}$ et $\boldsymbol{e}$ est A^{-1}

§ 4. Problème des vecteurs propres

1. Transformations linéaires. Au paragraphe précédent, on a défini la transformation linéaire comme une application linéaire de l'espace dans lui-même. Tous les résultats obtenus au § 3 pour les applications sont aussi vrais pour les transformations, à condition qu'on fasse quelques remarques importantes concernant l'expression analytique de la transformation.

A savoir, pour exprimer en coordonnées une application linéaire $\mathbf{A} : \mathscr{L} \to \tilde{\mathscr{L}}$, on choisit des bases dans les espaces $\mathscr{L}$ et $\tilde{\mathscr{L}}$. Si les espaces coïncident, il est tout naturel d'utiliser une même base pour les images et leurs antécédents. Aussi a-t-on la

DÉFINITION. On appelle *matrice de la transformation linéaire* $\mathbf{A} : \mathscr{L} \to \mathscr{L}$ par rapport à la base $\boldsymbol{e} = \|e_1, \ldots, e_n\|$ la matrice dont les colonnes sont des colonnes de coordonnées des vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_n)$ par rapport à la base $\boldsymbol{e}$.

En accord avec cette définition, la formule (6) du 3 pour la matrice d'une transformation prend la forme

$$A' = S^{-1}AS. \tag{1}$$

L'ensemble des matrices A' déduites de la matrice donnée A suivant la formule (1) pour des S différentes est plus étroit que l'ensemble des matrices déduites de A suivant la formule (6) du § 3 pour des matrices de passage S et P non liées entre elles. Il en découle certaines particularités de l'expression analytique des transformations dont la principale est la suivante : un ensemble plus étroit peut ne pas contenir de matrice de forme canonique (8), § 3, et le théorème 2 du § 3 ne se vérifie pas pour les transformations.

Il ne faut pas croire que c'est une conséquence accidentelle due à la définition « malchanceuse » de la matrice d'une transformation. En effet, la matrice d'une application définit cette application et, par suite, toutes les propriétés de l'application se retrouvent parmi celles de sa matrice. Les propriétés de l'application sont celles des propriétés de la matrice qui demeurent invariantes, c'est-à-dire qui ne varient pas lorsqu'on passe à un autre couple de bases. Les autres propriétés de la matrice dépendent non seulement de l'application considérée mais aussi du couple de bases choisies. Le théorème 2 du § 3 signifie en effet que pour les espaces $\mathcal{L}$ et $\tilde{\mathcal{L}}$, donnés, l'unique propriété invariante de l'application est son rang, vu que toutes les applications d'un même rang se déterminent par la même matrice dans le couple de bases convenablement choisies.

Les transformations linéaires présentent plus de propriétés invariantes que les applications linéaires, ce qui est dû au fait que l'image et son antécédent appartiennent à un même espace. Il devient possible de parler de la position d'une image par rapport à son antécédent, par exemple de discuter leur colinéarité, question qui n'a aucun sens pour une application de l'espace $\mathcal{L}$ dans un espace $\tilde{\mathcal{L}}$ différent de $\mathcal{L}$.

On étudiera exclusivement dans ce paragraphe les transformations linéaires et leurs propriétés que les applications linéaires ne possèdent pas en général.

2. Sous-espaces invariants. Soit un espace vectoriel $\mathcal{L}$ et une transformation linéaire **A** de cet espace.

DÉFINITION. Le sous-espace $\mathcal{L}'$ de l'espace $\mathcal{L}$ est dit *invariant* par **A** si l'image $\mathbf{A}(x)$ de tout vecteur x de $\mathcal{L}'$ est encore dans $\mathcal{L}'$.

On peut formuler cette définition différemment, en disant que $\mathcal{L}'$ est invariant si $\mathbf{A}(\mathcal{L}')$ est un sous-espace dans $\mathcal{L}'$.

EXEMPLE 1. Considérons l'espace géométrique ordinaire des points et une rotation **A** de cet espace autour d'un axe donné p de l'angle α. Un vecteur se transforme par rotation en vecteur et, partant, la rotation **A** engendre une transformation dans l'espace tridimensionnel des vecteurs. Cette transformation est évidemment linéaire. Les vecteurs x portés par l'axe p engendrent un sous-espace invariant car $\mathbf{A}(x) = x$. Les vecteurs perpendiculaires à l'axe p constituent un sous-espace invariant bidimensionnel vu

que l'image par rotation de tout vecteur perpendiculaire à l'axe p est encore perpendiculaire à cet axe.

EXEMPLE 2. Le sous-espace nul se transforme toujours en lui-même et, par suite, est invariant par toute transformation.

EXEMPLE 3. L'espace $\mathscr{L}$ considéré comme son sous-espace est un sous-espace invariant.

EXEMPLE 4. Tout sous-espace est invariant par les transformations identique et nulle.

EXEMPLE 5. Le noyau d'une transformation et l'ensemble de ses valeurs sont des sous-espaces invariants.

Soit **A** une transformation linéaire dans l'espace vectoriel $\mathscr{L}$ de dimension n et soit $\mathscr{L}'$ un sous-espace de dimension k invariant par **A**. Choisissons dans $\mathscr{L}$ une base $\|e_1, \ldots, e_n\|$ telle que les vecteurs $e_1, \ldots, e_k$ soient dans $\mathscr{L}'$. La matrice A de la transformation **A** peut être divisée en quatre blocs :

$$A = \left\| \begin{array}{c|c} A_1 & A_2 \\ \hline A_3 & A_4 \end{array} \right\|.$$

Les blocs A_1, A_2, A_3 et A_4 sont respectivement des matrices de types (k, k), $(k, n - k)$, $(n - k, k)$ et $(n - k, n - k)$. Démontrons que la matrice A_3 est nulle, autrement dit que les éléments α^i_j de la matrice A sont nuls pour les valeurs des indices $j = 1, \ldots, k$ et $i = k + 1, \ldots, n$. En effet, les k premières colonnes de la matrice A sont des colonnes de coordonnées des vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_k)$. Vu que $\mathscr{L}'$ est un sous-espace invariant, ces vecteurs se trouvent dans $\mathscr{L}'$, de sorte que leurs composantes α^i_j par rapport aux vecteurs de base $e_{k+1}, \ldots, e_n$ sont nulles d'après la proposition 3 du § 2.

Inversement, il est aisé de voir que s'il existe une base par rapport à laquelle la matrice de la transformation linéaire **A** prend la forme

$$\left\| \begin{array}{c|c} A_1 & A_2 \\ \hline O & A_4 \end{array} \right\|, \tag{2}$$

l'enveloppe linéaire des vecteurs $e_1, \ldots, e_k$ est un sous-espace invariant. En effet, il résulte de (2) que pour tous $j = 1, \ldots, k$

$$\mathbf{A}(e_j) = \sum_{i=1}^{k} \alpha^i_j e_i,$$

et, par suite, l'image de la combinaison linéaire des vecteurs $e_1, \ldots, e_k$ est une combinaison linéaire des mêmes vecteurs. Résumons ce qui vient d'être dit.

PROPOSITION 1. *La matrice A de la transformation* **A** *est de la forme* (2) *si*

et seulement si l'enveloppe linéaire des vecteurs $e_1, \ldots, e_k$ est un sous-espace invariant.

La transformation **A** fait correspondre à chaque vecteur du sous-espace invariant $\mathscr{L}'$ un vecteur de $\mathscr{L}'$, ce qui définit une transformation de l'espace $\mathscr{L}'$ qu'on appellera *restriction* de la transformation **A** au sous-espace $\mathscr{L}'$ et qu'on notera **A**′. Pour les vecteurs de $\mathscr{L}'$ on a $\mathbf{A}'(x) = \mathbf{A}(x)$, tandis que pour les vecteurs n'appartenant pas à $\mathscr{L}'$ la transformation **A**′ n'est pas définie. La transformation **A**′ ne diffère de **A** que par l'ensemble des vecteurs pour lesquels elle est définie.

La restriction de la transformation linéaire est assurément une transformation linéaire.

Conservons les notations utilisées pour la démonstration de la proposition précédente. Il n'est pas difficile de démontrer que dans la base $\| e_1, \ldots, e_k \|$ de l'espace $\mathscr{L}'$ le bloc A_1 de la matrice (2) est la matrice de la transformation **A**′.

3. Vecteurs propres. Considérons un sous-espace unidimensionnel $\mathscr{L}_1$ de l'espace vectoriel $\mathscr{L}$. La base dans $\mathscr{L}_1$ comprend un seul vecteur x non nul, de sorte que tout vecteur y de $\mathscr{L}_1$ est de la forme αx, où α est un nombre convenablement choisi. Si $\mathscr{L}_1$ est invariant par la transformation linéaire **A** définie sur $\mathscr{L}$, on a $\mathbf{A}(x) \in \mathscr{L}_1$. Il existe donc un nombre λ tel que

$$\mathbf{A}(x) = \lambda x. \tag{3}$$

Inversement, si la condition (3) est satisfaite pour un vecteur non nul de $\mathscr{L}_1$, elle l'est également pour tout vecteur de $\mathscr{L}_1$ (on le vérifie aisément en multipliant les deux membres de l'égalité (3) par un nombre arbitraire). Par suite, $\mathscr{L}_1$ est un sous-espace invariant.

DÉFINITION. On appelle *vecteur propre* de la transformation **A** un vecteur non nul x satisfaisant à la condition (3). Le nombre λ de l'égalité (3) est appelé *valeur propre*. On dit que le vecteur propre x est *associé* à la valeur propre λ.

On a vu que chaque sous-espace invariant unidimensionnel se définit par un vecteur propre et, inversement, chaque vecteur propre définit un sous-espace invariant unidimensionnel.

Proposons-nous de trouver tous les vecteurs propres de la transformation linéaire donnée **A**. Ce problème est d'une grande importance aussi bien pour les espaces de dimension finie que pour les espaces de dimension infinie. On l'étudiera pour les espaces de dimension finie n.

L'espace $\mathscr{L}$ étant rapporté à une base, l'égalité (3) s'écrit sous la forme de la relation $A\xi = \lambda\xi$ qui relie la matrice A de la transformation **A** et la colonne de coordonnées ξ du vecteur x. En notant E la matrice unité d'ordre n, on peut mettre cette relation sous la forme

$$(A - \lambda E)\xi = o \tag{4}$$

ou de façon détaillée

$$\left.\begin{aligned}(\alpha^1_1 - \lambda)\xi^1 + \alpha^1_2\xi^2 + \ldots + \alpha^1_n\xi^n = 0,\\ \alpha^2_1\xi^1 + (\alpha^2_2 - \lambda)\xi^2 + \ldots + \alpha^2_n\xi^n = 0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \alpha^n_1\xi^1 + \alpha^n_2\xi^2 + \ldots + (\alpha^n_n - \lambda)\xi^n = 0.\end{aligned}\right\} \qquad (5)$$

Traitons les égalités (5) comme un système d'équations permettant de trouver les composantes $\xi^1, \ldots, \xi^n$ du vecteur propre x. C'est un système de n équations linéaires homogènes à n inconnues. Le vecteur x devant être différent de zéro, on ne s'intéressera qu'aux solutions non triviales, autrement dit aux solutions dont l'une au moins des composantes ξ^i est non nulle. Le système admet une solution non triviale si son rang est strictement inférieur au nombre d'inconnues. Dans le cas considéré cela signifie que le déterminant de la matrice du système doit être nul :

$$\det(A - \lambda E) = \begin{vmatrix}\alpha^1_1 - \lambda\alpha^1_2 \ldots \lambda^1_n\\ \ldots\ldots\ldots\ldots\ldots\\ \alpha^n_1\alpha^n_2 \ldots \alpha^n_n - \lambda\end{vmatrix} = 0. \qquad (6)$$

L'égalité (6) est dénommée *équation caractéristique*. C'est la condition imposée au paramètre λ, à laquelle doivent satisfaire toutes les valeurs propres de la transformation **A**. Il va de soi que dans un espace réel aucune racine complexe de l'équation caractéristique n'est valeur propre, car l'égalité (3) perd alors son sens.

La dernière remarque étant prise en compte, la condition obtenue est aussi suffisante. En effet, le rang du système (5) diminue pour chaque valeur de λ satisfaisant à la condition (6), et ce système possède une solution non triviale. Cette solution est la colonne de coordonnées du vecteur propre x correspondant à un λ donné. On aboutit ainsi au théorème suivant.

THÉORÈME 1. *Dans un espace complexe, les seules racines de l'équation caractéristique sont des valeurs propres de la transformation. Il en est de même dans un espace réel pour les racines réelles de l'équation caractéristique.*

Le premier membre de l'équation caractéristique est un polynôme de degré n. En effet, selon la formule (6) du § 2, ch. V, le déterminant est égal à la somme algébrique de produits dont chacun renferme n éléments de la matrice. Les seuls éléments qui contiennent λ sont ceux de la diagonale principale. Il existe donc le produit

$$(\alpha^1_1 - \lambda)(\alpha^2_2 - \lambda)\ldots(\alpha^n_n - \lambda) \qquad (7)$$

dont chaque facteur renferme λ. Tout autre produit qui contient $\alpha^i_j (i \neq j)$

ne peut contenir les facteurs $(\alpha_i^i - \lambda)$ et $(\alpha_j^j - \lambda)$, de sorte que chaque terme de la somme, sauf (7), contient une puissance de λ dont l'exposant est au plus égal à $n - 2$. Chassons les parenthèses dans (7) et écrivons deux termes de plus haut degré en λ : $(-1)^n\lambda^n + (-1)^{n-1}(\alpha_1^1 + \ldots + \alpha_n^n)\lambda^{n-1}$. Ces termes sont aussi dominants dans tout le polynôme. Le terme constant du polynôme est égal à sa valeur pour $\lambda = 0$, soit à $\det(A - 0E) = \det A$. Ainsi donc, le premier membre de l'égalité (6) est le polynôme de la forme

$$(-1)^n\lambda^n + (-1)^{n-1}\lambda^{n-1} \sum_{i=1}^{n} \alpha_i^i + \ldots + \det A.$$

Ce polynôme est appelé *polynôme caractéristique de la matrice* A. Il n'est pas difficile d'expliciter les autres coefficients, mais on n'en a pas besoin. Comme on le sait, un polynôme de degré n ne peut avoir plus de n racines différentes et possède toujours au moins une racine complexe. Si l'on considère un espace réel, il peut arriver (pour un n pair) que l'équation caractéristique ne possède aucune racine réelle et, par suite, la transformation linéaire n'a ni valeurs propres ni vecteurs propres. A titre d'exemple, on peut citer la rotation du plan.

Dans un espace complexe, toute transformation linéaire présente au moins une valeur propre et, partant, au moins un vecteur propre.

4. Propriétés des vecteurs propres et des valeurs propres.

PROPOSITION 2. *Tous les vecteurs propres associés à une même valeur propre constituent avec le vecteur nul un sous-espace vectoriel.*

L'assertion découle immédiatement du fait que les colonnes de coordonnées de ces vecteurs constituent l'ensemble de toutes les solutions du système d'équations linéaires homogènes.

THÉORÈME 2. *Les vecteurs propres* $x_1, \ldots, x_k$ *sont linéairement indépendants s'ils sont associés à des valeurs propres deux à deux différentes.*

Démontrons ce théorème par récurrence. Vérifions l'assertion pour deux vecteurs propres x_1 et x_2 associés à des valeurs propres différentes λ_1 et λ_2. Supposons qu'ils sont linéairement dépendants. Les vecteurs x_1 et x_2 étant non nuls, il existe un nombre α tel que $x_1 = \alpha x_2$. En appliquant la transformation $\mathbf{A}$, on obtient $\mathbf{A}(x_1) = \lambda_1 x_1$ et

$$\mathbf{A}(x_1) = \alpha\mathbf{A}(x_2) = \alpha\lambda_2 x_2 = \lambda_2 x_1.$$

Cela signifie que $\lambda_1 x_1 = \lambda_2 x_1$, ce qui est impossible pour $\lambda_1 \neq \lambda_2$. Ainsi donc, les vecteurs x_1 et x_2 sont linéairement indépendants.

Admettons maintenant que tout système de $k - 1$ vecteurs propres associés à des valeurs propres différentes est libre et démontrons-le pour un système de k vecteurs.

Supposons que le système des vecteurs $x_1, \ldots, x_k$ satisfait à la condition du théorème. Considérons leur combinaison linéaire égale à zéro

$$\alpha_1 x_1 + \ldots + \alpha_k x_k = o. \tag{8}$$

Son image par la transformation **A** et son produit par λ_k sont respectivement

$$\alpha_1 \lambda_1 x_1 + \ldots + \alpha_k \lambda_k x_k = o$$

et

$$\alpha_1 \lambda_k x_1 + \ldots + \alpha_k \lambda_k x_k = o.$$

En retranchant ces égalités membre à membre, on obtient

$$\alpha_1(\lambda_1 - \lambda_k)x_1 + \ldots + \alpha_{k-1}(\lambda_{k-1} - \lambda_k)x_{k-1} = o. \tag{9}$$

Par hypothèse de récurrence, cette combinaison linéaire est triviale et par suite

$$\alpha_1(\lambda_1 - \lambda_k) = 0, \ldots, \alpha_{k-1}(\lambda_{k-1} - \lambda_k) = 0.$$

Puisque les valeurs propres sont deux à deux différentes, on a $\alpha_1 = \ldots \ldots = \alpha_{k-1} = 0$, si bien que l'égalité (8) entraîne $\alpha_k = 0$. Ainsi donc, la combinaison linéaire (8) est triviale. Le théorème est démontré.

Proposition 3. *Si A et A' sont deux matrices de la transformation* **A** *par rapport aux bases distinctes, les polynômes caractéristiques de ces matrices coïncident.*

En effet, selon la formule $A' = S^{-1}AS$ on a

$$\det(A' - \lambda E) = \det(S^{-1}AS - \lambda E) = \det S^{-1}(A - \lambda E)S = \\ = \det(A - \lambda E)\det S \det S^{-1} = \det(A - \lambda E).$$

Il découle de cette proposition qu'on peut appeler *polynôme caractéristique de la transformation* **A** le polynôme caractéristique de la matrice A.

Les coefficients du polynôme caractéristique sont des invariants liés à la transformation. En particulier, le déterminant de la matrice d'une transformation est indépendant du choix de la base. Un autre invariant important est le coefficient de $(-\lambda)^{n-1}$ appelé *trace* de la matrice : $\alpha_1^1 + \alpha_2^2 + \ldots \ldots + \alpha_n^n$.

Considérons maintenant un polynôme quelconque $P(\lambda) = \gamma_n\lambda^n + \ldots \ldots + \gamma_1\lambda + \gamma_0$. Si λ_0 est la racine de ce polynôme, $P(\lambda)$ est divisible par le binôme $\lambda - \lambda_0$, c'est-à-dire est égal au produit de $\lambda - \lambda_0$ par le polynôme $P_1(\lambda)$. Il peut arriver que $P(\lambda)$ est divisible non seulement par $\lambda - \lambda_0$ mais aussi par $(\lambda - \lambda_0)^s$ pour un entier $s > 1$, autrement dit, $P(\lambda)$ est de la forme $(\lambda - \lambda_0)^s P_2(\lambda)$, où P_2 est un polynôme. Le plus grand nombre s muni de cette propriété est appelé *multiplicité* de la racine λ_0. Les racines de multiplicité 1 sont dites *simples*.

THÉORÈME 3. *Si la valeur propre λ_0 de la transformation* **A** *est une racine de l'équation caractéristique de multiplicité s, on peut lui associer au plus s vecteurs propres linéairement indépendants.*

DÉMONSTRATION. Supposons qu'il existe k vecteurs propres linéairement indépendants associés à λ_0. Notons-les $e_1, \ldots, e_k$ et complétons-les par les vecteurs $e_{k+1}, \ldots, e_n$ jusqu'à la base de l'espace $\mathscr{L}$. Rapportée à cette base, la matrice A de la transformation **A** est de la forme

$$A = \left\| \begin{array}{c|c} \begin{matrix} \lambda_0 & & 0 \\ & \ddots & \\ 0 & & \lambda_0 \end{matrix} & \\ \cline{1-1} O & B \end{array} \right\|,$$

où B est une matrice à n lignes et $n-k$ colonnes. En effet, pour chaque $i \leqslant k$ la i-ième colonne est la colonne de coordonnées du vecteur $\mathbf{A}(e_i) = \lambda_0 e_i$, dont tous les éléments sont des zéros sauf le i-ième élément égal à λ_0. Considérons la matrice $A - \lambda E$ et développons son déterminant suivant chacune des k premières colonnes. Il vient $\det(A - \lambda E) = (\lambda_0 - \lambda)^k P(\lambda)$. D'après la définition de la multiplicité on voit maintenant que $k \leqslant s$.

Le lecteur peut vérifier à titre d'exercice que la transformation linéaire de l'espace bidimensionnel, définie par la matrice $\left\| \begin{matrix} 1 & 1 \\ 0 & 1 \end{matrix} \right\|$, possède une valeur propre de multiplicité 2 et un seul vecteur propre indépendant.

PROPOSITION 4. *La transformation linéaire admet une valeur propre égale à zéro si et seulement si elle n'est pas bijective.*

En effet, d'après la proposition 6 du § 3, la transformation est bijective si et seulement si son rang est n, autrement dit si le déterminant de sa matrice est différent de zéro. Or le déterminant est égal au terme constant du polynôme caractéristique, de sorte que le déterminant est nul si et seulement si le zéro est la racine de l'équation caractéristique.

Signalons que pour un vecteur propre associé à la valeur propre nulle, on a $\mathbf{A}(x) = o$. Cela signifie que le noyau de **A** renferme le vecteur non nul x.

PROPOSITION 5. *Soit* **A** *une transformation linéaire de l'espace vectoriel réel. A toute racine complexe du polynôme caractéristique de la transformation* **A** *correspond un sous-espace invariant bidimensionnel.*

Si le nombre $\alpha + i\beta$ vérifie l'équation $\det(A - \lambda E) = 0$, le système d'équations linéaires $[A - (\alpha + i\beta)E]\xi = o$ possède une solution non triviale. Cette solution est une matrice-

colonne complexe à n éléments que l'on peut écrire sous la forme $\eta + i\zeta$, où η et ζ sont des matrices-colonnes réelles. L'égalité de deux matrices-colonnes complexes $A\xi = \lambda\xi$ est équivalente à deux égalités entre les matrices-colonnes réelles, à savoir :

$$A\eta = \alpha\eta - \beta\zeta, \qquad A\zeta = \beta\eta + \alpha\zeta. \tag{10}$$

Les matrices-colonnes η et ζ sont réelles et on peut leur faire correspondre des vecteurs qu'on notera y et z. Les égalités (10) signifient que $\mathbf{A}(y) = \alpha y - \beta z$ et $\mathbf{A}(z) = \beta y + \alpha z$. Il en découle que l'enveloppe linéaire des vecteurs y et z est un sous-espace invariant. En effet,

$$\mathbf{A}(\mu y + \nu z) = \mu\mathbf{A}(y) + \nu\mathbf{A}(z) = (\mu\alpha + \nu\beta)y + (\nu\alpha - \mu\beta)z. \tag{11}$$

Il reste à démontrer que les vecteurs y et z sont linéairement indépendants. Supposons le contraire : il existe alors des nombres μ et ν ($\mu^2 + \nu^2 \neq 0$) tels que $(\mu y + \nu z) = o$, d'où en vertu de (11), $(\mu\alpha + \nu\beta)y + (\nu\alpha - \mu\beta)z = o$. Posons pour fixer les idées que $y \neq o$. Multiplions les deux égalités précédentes respectivement par $\nu\alpha - \mu\beta$ et par ν et retranchons l'une de l'autre. Il vient,

$$[\nu(\mu\alpha + \nu\beta) - \mu(\nu\alpha - \mu\beta)]y = o,$$

d'où $(\nu^2 + \mu^2)\beta = 0$, c'est-à-dire $\beta = 0$, ce qui contredit l'hypothèse que $\alpha + i\beta$ est une racine complexe.

Remarque. Si $\alpha + i\beta$ vérifie l'équation algébrique à coefficients réels, il en est de même, comme on le sait, de son conjugué complexe $\alpha - i\beta$. Dans le cas considéré, à deux racines conjuguées complexes correspond un même sous-espace invariant. Laissons au lecteur le soin de le démontrer.

5. Diagonalisation de la matrice d'une transformation. On dira que la matrice carrée A d'éléments α^i_j est de *forme diagonale* ou est *diagonale* si $\alpha^i_j = 0$ pour $i \neq j$, autrement dit si ne sont différents de zéro que les éléments α^i_i situés sur la diagonale principale.

PROPOSITION 6. *La matrice de la transformation linéaire* $\mathbf{A}$ *par rapport à la base* $\|e_1, \ldots, e_n\|$ *est de forme diagonale si et seulement si tous les vecteurs de base sont des vecteurs propres de la transformation.*

En effet, si le vecteur e_i est un vecteur propre, on a $\mathbf{A}(e_i) = \lambda_i e_i$ et, par suite, le i-ième élément de la colonne de coordonnées du vecteur $\mathbf{A}(e_i)$ est λ_i, les autres étant nuls. Il ne reste qu'à se rappeler que la i-ième colonne de la matrice de $\mathbf{A}$ est la colonne de coordonnées de $\mathbf{A}(e_i)$.

L'assertion inverse se démontre de façon analogue.

Il découle du théorème 2 une condition suffisante, simple mais importante, pour qu'il existe une base formée des vecteurs propres de la transformation.

PROPOSITION 7. *Si une transformation possède n valeurs propres deux à deux différentes, il existe une base formée des vecteurs propres de cette transformation.*

Si le polynôme caractéristique d'une transformation admet des racines multiples, il est possible que sa matrice ne possède de forme diagonale en aucune base. L'exemple d'une telle transformation démunie de base de vecteurs propres a été fourni p. 197.

Toutefois, il peut arriver qu'une transformation linéaire possède moins de n valeurs propres tout en ayant une base formée de ses vecteurs propres. En effet, on peut fixer une base et considérer une transformation définie par une matrice diagonale quelconque à éléments égaux sur la diagonale principale. Pour les transformations nulle et identique tout vecteur non nul est un vecteur propre, et dans toute base leurs matrices sont diagonales.

La proposition 7 peut prendre la forme suivante.

PROPOSITION 8. *Si toutes les racines du polynôme caractéristique de la matrice A sont différentes, il existe une matrice S de déterminant non nul, telle que la matrice $S^{-1}AS$ est diagonale. Si la matrice A est réelle et l'on veut que S le soit aussi, il faut que les racines du polynôme caractéristique soient réelles.*

CHAPITRE VII

ESPACES EUCLIDIENS ET UNITAIRES

§ 1. Espaces euclidiens

1. Produit scalaire. L'espace vectoriel introduit dans le chapitre précédent diffère essentiellement de l'ensemble des vecteurs de l'espace géométrique ordinaire par le fait que dans l'espace vectoriel on n'a pas défini les notions de longueur et d'angle des vecteurs. On produira ces définitions dans le présent chapitre.

Si dans le premier chapitre on a défini le produit scalaire à partir des notions de longueur et d'angle, ici il vaut mieux de procéder de façon inverse. On définira axiomatiquement l'opération de multiplication scalaire, ensuite, en s'appuyant sur le produit scalaire, on introduira les notions de longueur et d'angle des vecteurs.

Les résultats obtenus jusque-là concernaient aussi bien les espaces réels que complexes. La multiplication scalaire se définit dans ces deux cas de façon différente. Le présent paragraphe est consacré aux espaces réels.

DÉFINITION. On appelle *espace euclidien* un espace vectoriel réel muni de l'opération de multiplication scalaire qui à tout couple de vecteurs x et y associe un nombre réel (noté (x, y)), tout en vérifiant les conditions suivantes, quels que soient les vecteurs x, y et z et le nombre α :

1° $(x, y) = (y, x)$;
2° $(x + y, z) = (x, z) + (y, z)$;
3° $(\alpha x, y) = \alpha(x, y)$;
4° $(x, y) > 0$, si $x \neq o$.

Indiquons les corollaires simples des axiomes 1° à 4° :

1. $(x, \alpha y) = (\alpha y, x) = \alpha(y, x)$ et, par suite, on a toujours

$$(x, \alpha y) = \alpha(x, y). \tag{1}$$

2. De façon analogue on démontre l'identité

$$(x, y + z) = (x, y) + (x, z). \tag{2}$$

3. En appliquant successivement les axiomes 2° et 3° et les deux corollaires précédents, on démontre aisément que pour tous vecteurs et nombres,

$$\left.\begin{aligned}\left(\sum_{i=1}^{s}\alpha_i x_i, y\right) &= \sum_{i=1}^{s}\alpha_i (x_i, y),\\ \left(x, \sum_{k=1}^{p}\beta_k y_k\right) &= \sum_{k=1}^{p}\beta_k (x, y_k).\end{aligned}\right\}\qquad (3)$$

4. Quel que soit le vecteur x, on a

$$(x, o) = 0. \qquad (4)$$

En effet, posons $o = 0x$. Alors $(x, o) = 0(x, x) = 0$.

EXEMPLE 1. Pour les vecteurs (segments orientés) on a défini le produit scalaire comme produit de leurs longueurs par le cosinus de l'angle qu'ils forment. Le produit scalaire ainsi défini possède les propriétés 1° à 4° et dépend du choix de l'unité de longueur. Par conséquent, si l'unité de longueur est choisie, les vecteurs de l'espace géométrique ordinaire forment par définition un espace euclidien tridimensionnel.

EXEMPLE 2. Dans l'espace vectoriel des matrices-colonnes à n éléments on peut introduire le produit scalaire en faisant correspondre à tout couple de matrices-colonnes $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ le nombre

$$\xi^1\eta^1 + \xi^2\eta^2 + \ldots + \xi^n\eta^n, \qquad (5)$$

où ξ^i et η^i sont les éléments des matrices-colonnes.

Le nombre (5) est le produit matriciel de la matrice-ligne $\boldsymbol{\xi}$ par la matrice-colonne $\boldsymbol{\eta}$. Le lecteur s'assurera aisément que les axiomes 1° à 4° sont vérifiés. Muni du produit scalaire ainsi défini, l'espace des matrices-colonnes à n éléments devient un espace euclidien de dimension n.

EXEMPLE 3. Dans l'espace des fonctions continues sur le segment [0, 1] on peut introduire le produit scalaire suivant la formule

$$(f, g) = \int_0^1 f(\xi)g(\xi)d\xi.$$

Les axiomes 1° à 4° découlent des propriétés bien connues de l'intégrale définie.

2. Longueur et angle. Dans le chapitre I, on a obtenu les formules exprimant la longueur du vecteur et l'angle des vecteurs au moyen du produit scalaire. Conformément à ces formules introduisons la

DÉFINITION. On appelle *longueur* du vecteur x et on la note $|x|$ le nombre $\sqrt{(x, x)}$. On appelle *angle* des vecteurs x et y tout nombre φ satisfaisant

à la condition

$$\cos\varphi = \frac{(x, y)}{|x|\,|y|} . \tag{6}$$

En vertu de l'axiome 4°, la longueur du vecteur est un nombre réel positif (ia racine est prise dans son sens arithmétique). La longueur du vecteur est nulle si et seulement si ce vecteur est nul.

Quant à la définition de l'angle des vecteurs, la situation se complique. En effet, il nous faut démontrer que la valeur absolue du second membre de l'égalité (6) est au plus égale à l'unité. Or ce fait résulte de l'inégalité

$$(x, y)^2 \leqslant (x, x)(y, y) \tag{7}$$

liée aux noms de Schwarz, Cauchy et Bouniakovski. Démontrons cette inégalité.

Soient x et y des vecteurs quelconques de l'espace euclidien. Pour tous α et β on a la relation

$$(\alpha x + \beta y, \alpha x + \beta y) = \alpha^2(x, x) + 2\alpha\beta(x, y) + \beta^2(y, y) \geqslant 0, \tag{8}$$

cette relation devenant nulle si et seulement si $\alpha x + \beta y = o$. En posant ici $\alpha = (y, y)$ et $\beta = -(x, y)$, on obtient

$$(y, y)[(x, x)(y, y) - (x, y)^2] \geqslant 0,$$

d'où il découle l'inégalité cherchée si $y \neq o$. Dans le cas de $y = o$, la relation (7) est évidente.

L'égalité a lieu dans la formule (7) si et seulement si x et y sont linéairement dépendants. En effet, s'ils sont indépendants, on a l'inégalité stricte dans (8) pour $\alpha^2 + \beta^2 \neq 0$. Donc, l'inégalité stricte est aussi vérifiée dans (7). Inversement, soit une combinaison linéaire non triviale $\alpha x + \beta y = 0$. En la multipliant scalairement par x puis par y, on obtient les égalités

$$\alpha(x, x) + \beta(y, x) = 0 \quad \text{et} \quad \alpha(x, y) + \beta(y, y) = 0.$$

Le déterminant d'un système homogène ayant une solution non triviale est égal à zéro, d'où vient l'égalité cherchée.

L'inégalité (7) entraîne encore une inégalité simple et utile, à savoir

$$|x + y| \leqslant |x| + |y|. \tag{9}$$

Elle découle de la suite de relations suivante:

$$(x + y, x + y) = |x|^2 + 2(x, y) + |y|^2 \leqslant |x|^2 + 2|x|\,|y| + |y|^2 = (|x| + |y|)^2.$$

L'égalité s'établit si $(x, y) = |x|\,|y|$, c'est-à-dire si et seulement si l'angle entre x et y est nul. L'inégalité (9) s'appelle *inégalité triangulaire* car, si les vecteurs sont des segments orientés, elle signifie qu'un côté du triangle est strictement inférieur à la somme de ses deux autres côtés.

On invite le lecteur d'écrire les inégalités (7) et (9) pour les espaces euclidiens considérés dans les exemples 2 et 3.

Les vecteurs x et y sont dits *perpendiculaires* ou *orthogonaux* si $(x, y) = 0$. Si au moins un des vecteurs x et y est nul, le second membre de (6) perd son sens. On admet par définition que le vecteur nul est orthogonal à tout vecteur.

PROPOSITION 1. *Seul le vecteur nul est orthogonal à tout vecteur.*

En effet, soit $(x, y) = 0$ pour tout y. En posant $y = x$, on a $(x, x) = 0$, ce qui n'est possible que pour le vecteur nul.

3. Base orthonormée. Le système des vecteurs $f_1, \ldots, f_m$ d'un espace euclidien est dit *orthonormé* si $(f_i, f_j) = 0$ pour $i \neq j$ et $(f_i, f_i) = 1$, quels que soient i et j.

PROPOSITION 2. *Tout système orthonormé de vecteurs est libre.*

DÉMONSTRATION. Soit $\{f_1, \ldots, f_m\}$ un système orthonormé de vecteurs. Considérons l'égalité $\alpha_1 f_1 + \ldots + \alpha_m f_m = o$. Il s'ensuit que $\alpha_i = 0$ pour tout i. En effet, multiplions scalairement par f_i les deux membres de l'égalité. Tous les termes, sauf le i-ième s'annulent et l'on obtient $\alpha_i (f_i, f_i) = = \alpha_i = 0$. Ainsi, toute combinaison linéaire des vecteurs $f_1, \ldots, f_m$ égale à zéro est nécessairement triviale. La proposition est démontrée.

THÉORÈME 1. *Dans un espace euclidien de dimension n il existe un système orthonormé de n vecteurs.*

Signalons qu'en vertu de la proposition 2, ce système de vecteurs est une base. Une telle base est dite *orthonormée*.

La démonstration sera faite par récurrence.

1) Pour $n = 1$, l'assertion est évidente. Si f est un vecteur non nul, le vecteur $e = |f|^{-1} f$ forme un système orthonormé à un vecteur.

2) Supposons que dans tout espace euclidien de dimension $n - 1$ il existe une base orthonormée et démontrons qu'il en est de même pour un espace euclidien quelconque $\mathscr{E}_n$ de dimension n. Soit $\|f_1, \ldots, f_n\|$ une base quelconque dans $\mathscr{E}_n$. L'enveloppe linéaire des vecteurs $f_1, \ldots, f_{n-1}$ est un espace euclidien de dimension $n - 1$, de sorte que par hypothèse de récurrence, il y existe un système orthonormé de $n - 1$ vecteurs $e_1, \ldots, e_{n-1}$. Considérons le vecteur $g_n = f_n - \alpha_1 e_1 - \ldots - \alpha_{n-1} e_{n-1}$. Choisissons les coefficients $\alpha_1, \ldots, \alpha_{n-1}$ de manière que le vecteur g_n soit orthogonal à tous les vecteurs $e_1, \ldots, e_{n-1}$. Etant donné que le système $\{e_1, \ldots, e_{n-1}\}$ est orthonormé, on a $(g_n, e_i) = (f_n, e_i) - \alpha_i$, d'où $\alpha_i = (f_n, e_i)$ pour tous les $i = 1, \ldots, n - 1$. Considérons maintenant le vecteur $e_n = |g_n|^{-1} g_n$. Sa longueur est égale à l'unité et il est orthogonal aux vecteurs $e_1, \ldots, e_{n-1}$. On voit aussitôt que le système $\{e_1, \ldots, e_n\}$ est orthonormé.

La méthode utilisée pour la démonstration du théorème est appelée

méthode d'orthogonalisation. Pour s'en servir, on commence par normer le vecteur f_1, c'est-à-dire qu'on construit le vecteur $e_1 = |f_1|^{-1}f_1$. Ensuite on recherche la base orthonormée $\|e_1, e_2\|$ dans l'enveloppe linéaire des vecteurs f_1, f_2. Cela se fait de la même façon que dans le point 2 de la démonstration. Après quoi, on construit de la même manière la base orthonormée dans l'enveloppe linéaire des vecteurs f_1, f_2, f_3, etc.

4. Expression du produit scalaire en fonction de composantes des facteurs. Soit l'espace euclidien $\mathscr{E}_n$ rapporté à une base donnée $\|e_1, \ldots, e_n\|$. Tous vecteurs x et y se mettent alors sous la forme $x = \sum_i \xi^i e_i$ et $y = \sum_j \eta^j e_j$. Ici et plus loin, si le contraire n'est pas spécifié, l'indice de sommation prend les valeurs allant de l'unité à la dimension n de l'espace. On a donc

$$(x, y) = \left(\sum_i \xi^i e_i, \sum_j \eta^j e_j\right).$$

En utilisant les formules (3), on peut écrire le produit scalaire sous la forme

$$(x, y) = \sum_{i,j} \xi^i \eta^j (e_i, e_j). \tag{10}$$

Si la base est orthonormée, on a $(e_i, e_j) = 0$ pour $i \neq j$, et la somme ne contient que les termes pour lesquels $i = j$. Puisque $(e_i, e_j) = 1$, dans une base orthonormée on a

$$(x, y) = \sum_i \xi^i \eta^i. \tag{11}$$

Etant donné une base quelconque, considérons tous les produits scalaires possibles (e_i, e_j) $(i, j = 1, \ldots, n)$ que l'on note habituellement g_{ij} et l'on écrit sous la forme d'une matrice carrée :

$$\Gamma = \begin{Vmatrix} g_{11} & \ldots & g_{1n} \\ \ldots & \ldots & \ldots \\ g_{n1} & \ldots & g_{nn} \end{Vmatrix} = \begin{Vmatrix} (e_1, e_1) & \ldots & (e_1, e_n) \\ \ldots & \ldots & \ldots \\ (e_n, e_1) & \ldots & (e_n, e_n) \end{Vmatrix}. \tag{12}$$

La matrice (12) est appelée *matrice de Gram* de la base $\|e_1, \ldots, e_n\|$. En vertu de la commutativité de la multiplication scalaire, on a $g_{ij} = g_{ji}$ et par suite, la matrice satisfait à la condition ${}^t\Gamma = \Gamma$, c'est-à-dire ne varie pas par transposition. Une telle matrice est dite *symétrique*.

Désignons par $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ les colonnes de coordonnées des vecteurs x et y. On vérifie aisément en multipliant les matrices que l'égalité (10) peut être écrite

sous la forme matricielle suivante :

$$(x, y) = \boldsymbol{\xi}\Gamma\boldsymbol{\eta}. \tag{13}$$

Une base est orthonormée si et seulement si sa matrice de Gram Γ est une matrice unité, de sorte qu'on a pour une base orthonormée

$$(x, y) = \boldsymbol{\xi}\boldsymbol{\eta}, \tag{14}$$

ce qui coïncide avec (11).

5. Lien entre les matrices de Gram des bases différentes. Soient données deux bases $\|e_1, \ldots, e_n\|$ et $\|e_1', \ldots, e_n'\|$ liées par la matrice de passage S suivant les formules $e_i' = \sum_k \sigma_i^k e_k$, où par σ_i^k sont désignés les éléments de S.

Pour des i et j quelconques on a maintenant en vertu de la formule (10)

$$g_{ij}' = (e_i', e_j') = \left(\sum_k \sigma_i^k e_k, \sum_l \sigma_j^l e_l\right) = \sum_{k,l} \sigma_i^k \sigma_j^l g_{kl}.$$

L'égalité obtenue exprime un élément de la matrice de Gram Γ' de la base e' au moyen des éléments de la matrice de Gram de la base e. L'ensemble de ces égalités pour tous les i et j est équivalent à la relation matricielle

$$\Gamma' = {}^tS\Gamma S. \tag{15}$$

On le vérifie aisément en explicitant le second membre de (15).

Considérons la formule (15) dans le cas particulier où la base e est orthonormée. Alors $\Gamma = E$ et $\Gamma' = {}^tSS$. En calculant le déterminant des deux membres de l'égalité, on obtient $\det \Gamma' = \det {}^tS \det S = (\det S)^2$. Vu que la base e' est arbitraire, on a la

PROPOSITION 3. *Le déterminant de la matrice de Gram de toute base est strictement positif.*

Cette proposition peut être renforcée de la façon suivante.

THÉORÈME 2. *Soient $x_1, \ldots, x_k$ des vecteurs arbitraires (non nécessairement linéairement indépendants) dans l'espace euclidien. Le déterminant de la matrice*

$$\begin{Vmatrix} (x_1, x_1) & \ldots & (x_1, x_k) \\ \ldots & \ldots & \ldots \\ (x_k, x_1) & \ldots & (x_k, x_k) \end{Vmatrix}$$

composée de leurs produits scalaires est strictement positif si les vecteurs sont linéairement indépendants, et égal à zéro s'ils sont linéairement dépendants.

La première assertion du théorème se déduit immédiatement de la proposition 3, car si $x_1, \ldots, x_k$ sont linéairement indépendants, ils constituent une base dans leur enveloppe linéaire.

Démontrons la seconde assertion. Si les vecteurs sont linéairement dépendants, on a l'égalité $\alpha_1 x_1 + \ldots + \alpha_k x_k = o$, où tous les coefficients $\alpha_1, \ldots, \alpha_k$ ne sont pas nuls. En multipliant cette égalité scalairement par chacun des vecteurs $x_1, \ldots, x_k$, on aboutit au système d'équations linéaires

$$\begin{gathered}\alpha_1(x_1, x_1) + \ldots + \alpha_k(x_1, x_k) = 0,\\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\ \alpha_1(x_k, x_1) + \ldots + \alpha_k(x_k, x_k) = 0,\end{gathered}$$

vérifié par les coefficients $\alpha_1, \ldots, \alpha_k$. Le système ayant une solution non triviale, le déterminant de sa matrice est égal à zéro, ce qu'il fallait démontrer.

Signalons que l'inégalité de Cauchy-Bouniakovski démontrée plus haut représente un cas particulier de ce théorème pour $k = 2$.

6. Matrices orthogonales. Supposons que dans la formule (15), $\Gamma = \Gamma' = E$, autrement dit admettons que les deux bases sont orthonormées. La formule prend alors la forme

$$'SS = E. \tag{16}$$

Définition. On dit que la matrice est *orthogonale* si elle satisfait à la condition (16).

On voit que les seules matrices orthogonales peuvent servir de matrices de passage d'une base orthonormée à l'autre. L'égalité (16) est équivalente à la suivante

$$'S = S^{-1}. \tag{17}$$

En vertu des propriétés de la matrice inverse, on a

$$S'S = E. \tag{18}$$

Cela signifie que la matrice $'S$ est également orthogonale.

En désignant les éléments de la matrice S par σ^i_k, on peut écrire les égalités (16) et (18) respectivement ainsi :

$$\sum_{k=1}^{n} \sigma^k_i \sigma^k_j = \begin{cases} 0, & i \neq j,\\ 1, & i = j,\end{cases} \tag{19}$$

$$\sum_{k=1}^{n} \sigma^i_k \sigma^j_k = \begin{cases} 0, & i \neq j,\\ 1, & i = j.\end{cases} \tag{20}$$

La relation (19) peut d'ailleurs être obtenue directement par la formule (11) si l'on se souvient que les colonnes de la matrice de passage sont les colonnes de coordonnées des nouveaux vecteurs de base par rapport à l'ancienne base.

En calculant le déterminant de chaque membre de l'égalité (16), on obtient $(\det S)^2 = 1$. Aussi le déterminant de la matrice orthogonale est-il égal à $+1$ ou -1.

On invite le lecteur à vérifier toutes les relations de ce point pour la matrice

$$\begin{Vmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{Vmatrix}.$$

7. Supplémentaire orthogonal d'un sous-espace. Considérons un sous-espace $\mathscr{E}_k$ de dimension k dans un espace euclidien $\mathscr{E}_n$ de dimension n.

DÉFINITION. On appelle *supplémentaire orthogonal* du sous-espace $\mathscr{E}_k$ l'ensemble de tous les vecteurs orthogonaux à chacun des vecteurs de $\mathscr{E}_k$.

Le supplémentaire orthogonal du sous-espace $\mathscr{E}_k$ est noté $\mathscr{E}_k^{\perp}$.

PROPOSITION 4. *Le supplémentaire orthogonal du sous-espace $\mathscr{E}_k$ est un sous-espace de dimension $n - k$.*

DÉMONSTRATION. Soit $\|a_1, \ldots, a_k\|$ une base dans $\mathscr{E}_k$. Le vecteur x appartient à $\mathscr{E}_k^{\perp}$ si et seulement si

$$(x, a_1) = 0, \ldots, (x, a_k) = 0. \tag{21}$$

En effet, si x appartient à $\mathscr{E}_k^{\perp}$, les conditions (21) sont évidemment satisfaites. Inversement, si ces conditions sont satisfaites, x est orthogonal à tout vecteur a de $\mathscr{E}_k$ puisque

$$(x, a) = \left(x, \sum_{p=1}^{k} \lambda^p a_p\right) = \sum_{p=1}^{k} \lambda^p (x, a_p) = 0.$$

Choisissons dans $\mathscr{E}_n$ une base orthonormée et désignons par $\alpha_p^1, \ldots, \alpha_p^n$ les composantes du vecteur a_p (pour tout $p = 1, \ldots, k$) et par $\xi^1, \ldots, \xi^n$ les composantes du vecteur x. Les conditions (21) s'écrivent alors sous la forme du système homogène de k équations linéaires à n inconnues :

$$\alpha_1^1 \xi^1 + \ldots + \alpha_1^n \xi^n = 0,$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\alpha_k^1 \xi^1 + \ldots + \alpha_k^n \xi^n = 0.$$

La matrice du système est de rang k puisque ses lignes sont linéairement indépendantes en tant que matrices-lignes des composantes des vecteurs $a_1, \ldots, a_k$. L'ensemble de toutes les solutions du système définit, comme on l'a montré, $\mathscr{E}_k^{\perp}$. On connaît par ailleurs que l'ensemble de toutes les solutions de ce système définit un sous-espace de dimension $n - k$. La proposition est démontrée.

Considérons maintenant le supplémentaire orthogonal $(\mathscr{E}_k^{\perp})^{\perp}$ du sous-espace $\mathscr{E}_k^{\perp}$. Il découle de la définition que $\mathscr{E}_k \subseteq (\mathscr{E}_k^{\perp})^{\perp}$. Or, la dimension

de $(\mathscr{E}_k^{\perp})^{\perp}$ est $n - (n - k) = k$. Donc, selon la proposition 2 du § 2, ch. VI, on a $(\mathscr{E}_k^{\perp})^{\perp} = \mathscr{E}_k$.

Il est évident que $\mathscr{E}_k$ et $\mathscr{E}_k^{\perp}$ n'ont pas de vecteurs communs non nuls, d'où la

PROPOSITION 5. *L'espace euclidien $\mathscr{E}_n$ est la somme directe de l'un quelconque de ses sous-espaces $\mathscr{E}_k$ et du supplémentaire orthogonal de ce dernier.*

Ainsi, selon la proposition 5 du § 2, ch. VI, tout vecteur x de $\mathscr{E}_n$ se décompose de façon univoque en somme des vecteurs x' de $\mathscr{E}_k$ et x'' de $\mathscr{E}_k^{\perp}$. Le vecteur x' est appelé *projection orthogonale* de x sur $\mathscr{E}_k$. On voit immédiatement que le vecteur x'' est la projection orthogonale de x sur $\mathscr{E}_k^{\perp}$. La longueur de x'' est appelée *distance du vecteur x au sous-espace* $\mathscr{E}_k$. Elle possède la propriété suivante de minimalité.

PROPOSITION 6. *Si le vecteur x représente la somme des vecteurs x' de $\mathscr{E}_k$ et x'' de $\mathscr{E}_k^{\perp}$, tout vecteur y de $\mathscr{E}_k$, différent de x', vérifie la relation*

$$|x''| = |x - x'| < |x - y|.$$

DEMONSTRATION. En désignant $x' - y$ par z, on obtient $|x - y|^2 = |x' - y + x''|^2 = |z + x''|^2 = (z + x'', z + x'') = |z|^2 + 2(z, x'') + |x''|^2$. Mais $(z, x'') = 0$, vu que z se trouve dans $\mathscr{E}_k$ et, par suite, $|x - y|^2 = |x''|^2 + |z|^2$. Il s'ensuit immédiatement l'assertion nécessaire.

§ 2. Transformations linéaires dans l'espace euclidien

1. Transformation adjointe. Tout ce qui a été dit au chapitre précédent sur les transformations linéaires dans les espaces vectoriels s'applique également aux espaces euclidiens. Le produit scalaire introduit dans l'espace euclidien permet de définir des classes importantes de transformations que nous allons étudier dans ce paragraphe. L'exposé qui suivra se rapporte exclusivement aux espaces euclidiens réels.

DÉFINITION. La transformation linéaire $\mathbf{A}^*$ de l'espace euclidien est dite *adjointe* de la transformation donnée $\mathbf{A}$ si pour tous vecteurs x et y on a l'égalité

$$(\mathbf{A}(x), y) = (x, \mathbf{A}^*(y)). \tag{1}$$

Admettons que la transformation donnée $\mathbf{A}$ possède une adjointe $\mathbf{A}^*$. Cherchons comment sont liées les matrices des transformations $\mathbf{A}$ et $\mathbf{A}^*$ dans une base $\boldsymbol{e}$. Désignons les matrices de ces transformations respectivement par A et A^*, et les colonnes de coordonnées des vecteurs x et y par $\boldsymbol{\xi}$

et $\boldsymbol{\eta}$. L'égalité (1) s'écrit alors ainsi

$$^t(A\boldsymbol{\xi})\Gamma\boldsymbol{\eta} = {}^t\boldsymbol{\xi}\Gamma A^*\boldsymbol{\eta},$$

où Γ est la matrice de Gram de la base $\boldsymbol{e}$. Après des transformations évidentes, on obtient

$$^t\boldsymbol{\xi}\,({}^tA\Gamma - \Gamma A^*)\boldsymbol{\eta} = 0. \tag{2}$$

Vu que $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ sont des matrices-colonnes arbitraires, on peut en conclure que

$$^tA\Gamma - \Gamma A^* = O, \tag{3}$$

où O est la matrice nulle. Pour aboutir à cette conclusion, rappelons l'exemple 4 de la p. 160. On y a vu que pour toute matrice P et les colonnes $\boldsymbol{e}_i$ et $\boldsymbol{e}_j$ de la matrice unité le produit $^t\boldsymbol{e}_iP\boldsymbol{e}_j$ est égal à l'élément p_{ij} de la matrice P. En substituant à $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ les colonnes de la matrice unité, on peut montrer que tout élément de la matrice $^tA\Gamma - \Gamma A^*$ est égal à zéro.

Ainsi donc, les matrices des transformations $\mathbf{A}$ et $\mathbf{A}^*$ sont liées par la relation (3). En particulier, si la base est orthonormée et $\Gamma = E$, on a

$$A^* = {}^tA. \tag{4}$$

PROPOSITION 1. *Toute transformation linéaire de l'espace euclidien possède une transformation adjointe et une seule.*

Pour le démontrer, choisissons une base orthonormée et considérons la transformation $\mathbf{B}$ de matrice ^tA, si A est la matrice de la transformation donnée $\mathbf{A}$. Pour la transformation $\mathbf{B}$, la condition (1) est de toute évidence équivalente à $^t(A\boldsymbol{\xi})\boldsymbol{\eta} = {}^t\boldsymbol{\xi}({}^tA\boldsymbol{\eta})$. Donc, $\mathbf{B}$ est la transformation adjointe de $\mathbf{A}$. S'il y avait deux transformations adjointes $\mathbf{A}$, leurs matrices coïncideraient en vertu de (4). La proposition est démontrée.

Puisque $^t({}^tA) = A$, la formule (4) entraîne que

$$(\mathbf{A}^*)^* = \mathbf{A}. \tag{5}$$

2. Transformations symétriques. DÉFINITION. La transformation linéaire $\mathbf{A}$ de l'espace euclidien est dite *autoadjointe* ou *symétrique* si $\mathbf{A} = \mathbf{A}^*$.

Il découle immédiatement de la formule (4) la

PROPOSITION 2. *La transformation est symétrique si et seulement si sa matrice est symétrique (c'est-à-dire satisfait à la condition $A = {}^tA$) dans toute base orthonormée.*

Les valeurs propres et les vecteurs propres des transformations symétriques possèdent une série de propriétés importantes et intéressantes à l'exposé desquelles on va passer.

THÉORÈME 1. *Toutes les racines du polynôme caractéristique d'une transformation symétrique sont réelles.*

DÉMONSTRATION. Notons A la matrice de la transformation symétrique considérée dans une base orthonormée quelconque. Supposons que l'équation caractéristique $\det(A - \lambda E) = 0$ possède une racine complexe λ_0. Considérons le système d'équations linéaires

$$(A - \lambda_0 E)\boldsymbol{\xi} = o \tag{6}$$

à n inconnues $\xi^1, \ldots, \xi^n$ (n étant la dimension de l'espace). La matrice du système est complexe, de sorte que la solution $\boldsymbol{\xi}$ est en général une matrice-colonne complexe. Puisque $\det(A - \lambda_0 E) = 0$, il existe obligatoirement une solution non triviale. Soit $\boldsymbol{\xi}_0$ une solution non triviale. Portons $\boldsymbol{\xi}_0$ dans le système et multiplions à gauche les deux membres de l'égalité obtenue par la matrice-ligne ${}^t\bar{\boldsymbol{\xi}}_0$:

$${}^t\bar{\boldsymbol{\xi}}_0 A \boldsymbol{\xi}_0 = \lambda_0 {}^t\bar{\boldsymbol{\xi}}_0 \boldsymbol{\xi}_0. \tag{7}$$

Vu que ${}^t\bar{\boldsymbol{\xi}}_0 \boldsymbol{\xi}_0 = \xi_0^1 \bar{\xi}_0^1 + \ldots + \xi_0^n \bar{\xi}_0^n$ est un nombre réel strictement positif, pour aboutir à une contradiction il suffit de montrer que le nombre ${}^t\bar{\boldsymbol{\xi}}_0 A \boldsymbol{\xi}_0$ est réel. En effet, posons $\omega = {}^t\bar{\boldsymbol{\xi}}_0 A \boldsymbol{\xi}_0$. La matrice carrée d'ordre 1 ne varie pas par transposition, de sorte que

$$\omega = {}^t\omega = {}^t({}^t\bar{\boldsymbol{\xi}}_0 A \boldsymbol{\xi}_0) = {}^t\boldsymbol{\xi}_0 {}^tA \bar{\boldsymbol{\xi}}_0;$$

d'autre part, on a

$$\bar{\omega} = {}^t\boldsymbol{\xi}_0 \bar{A} \bar{\boldsymbol{\xi}}_0.$$

Or A est une matrice symétrique réelle et, par suite, $A = \bar{A} = {}^tA$. On a donc $\omega = \bar{\omega}$, d'où ω est réel. En divisant les deux membres de l'égalité (7) par le nombre non nul ${}^t\bar{\boldsymbol{\xi}}_0 \boldsymbol{\xi}_0$, on constate que λ_0 est obligatoirement réel.

L'assertion démontrée peut être formulée en termes de matrices.

PROPOSITION 3. *Si A est une matrice symétrique réelle, toutes les racines de l'équation* $\det(A - \lambda E) = 0$ *sont réelles.*

THÉORÈME 2. *Les vecteurs propres d'une transformation symétrique* **A**, *associés à différentes valeurs propres sont orthogonaux.*

En effet, soit $\lambda \neq \mu$ et $\mathbf{A}(x) = \lambda x$, $\mathbf{A}(y) = \mu y$. Alors

$$(\mathbf{A}(x), y) = \lambda(y, y).$$

Or on peut obtenir d'une autre façon

$$(\mathbf{A}(x), y) = (x, \mathbf{A}(y)) = \mu(x, y).$$

Il ressort de ces deux égalités que $(\lambda - \mu)(x, y) = 0$, d'où $(x, y) = 0$, ce qu'il fallait démontrer.

THÉORÈME 3. *Si le sous-espace $\mathscr{E}'$ est invariant par la transformation symétrique* **A**, *il en est de même de son supplémentaire orthogonal* $(\mathscr{E}')^\perp$.

DÉMONSTRATION. On a par hypothèse que l'image $\mathbf{A}(x)$ de tout x de $\mathscr{E}'$ appartient à $\mathscr{E}'$. Cela signifie que $(\mathbf{A}(x), y) = 0$ pour tout y de $(\mathscr{E}')^{\perp}$. La transformation $\mathbf{A}$ étant symétrique, on en déduit que $(x, \mathbf{A}(y)) = 0$ et, partant, $\mathbf{A}(y)$ appartient à $(\mathscr{E}')^{\perp}$, ce qui démontre le théorème.

On est maintenant en mesure de démontrer le théorème qui permet de décrire toutes les transformations symétriques possibles. On l'appellera *théorème fondamental des transformations symétriques*.

THÉORÈME 4. *Soit* $\mathbf{A}$ *une transformation linéaire symétrique dans l'espace euclidien* $\mathscr{E}_n$ *de dimension n. Il existe alors dans* $\mathscr{E}_n$ *une base orthonormée des vecteurs propres de la transformation* $\mathbf{A}$.

Raisonnons par récurrence sur la dimension n de l'espace $\mathscr{E}_n$. Pour un espace unidimensionnel $\mathscr{E}_1$ le théorème est évident, car chaque vecteur dans cet espace est un vecteur propre de $\mathbf{A}$, de sorte qu'il suffit de prendre pour base cherchée tout vecteur de longueur 1.

Supposons maintenant que le théorème est démontré pour un espace de dimension $k - 1$ et démontrons-le pour un espace de dimension k. Selon le théorème 1, la transformation symétrique $\mathbf{A}$ de $\mathscr{E}_k$ a au moins une valeur propre *) et, par suite, au moins un sous-espace invariant unidimensionnel. Désignons ce sous-espace par $\mathscr{E}_1$ et son vecteur unitaire par e. En vertu du théorème 3, le supplémentaire orthogonal $\mathscr{E}_{k-1}$ du sous-espace $\mathscr{E}_1$ est un sous-espace de dimension $k - 1$, également invariant par $\mathbf{A}$.

Considérons la restriction $\mathbf{A}'$ de la transformation $\mathbf{A}$ au sous-espace $\mathscr{E}_{k-1}$ (voir p. 193). On constate aisément que c'est une transformation symétrique dans $\mathscr{E}_{k-1}$. En effet, l'égalité $(\mathbf{A}(x), y) = (x, \mathbf{A}(y))$ est vérifiée par tous les vecteurs de $\mathscr{E}_k$ et, par suite, par tous les vecteurs de $\mathscr{E}_{k-1}$, pour lesquels d'ailleurs on a, par définiton, $\mathbf{A}'(x) = \mathbf{A}(x)$. Si x est un vecteur propre de la transformation $\mathbf{A}'$, il est aussi un vecteur propre de $\mathbf{A}$ vu que $\mathbf{A}'(x) = \mathbf{A}(x) = \lambda x$.

Par hypothèse de récurrence, il existe dans $\mathscr{E}_{k-1}$ une base orthonormée $\|e_1, \ldots, e_{k-1}\|$ des vecteurs propres de la transformation $\mathbf{A}'$. Considérons le système des vecteurs $e_1, \ldots, e_{k-1}, e$. Tous ces vecteurs sont deux à deux orthogonaux : $e_1, \ldots, e_{k-1}$ par construction, et e est orthogonal à chacun d'eux car $\mathscr{E}_{k-1}$ est le supplémentaire orthogonal de $\mathscr{E}_1$. La longueur de chacun des vecteurs considérés est 1. Chacun d'eux est un vecteur propre de la transformation $\mathbf{A}$. Ainsi donc, le système des vecteurs $e_1, \ldots, e_{k-1}, e$ est la base qu'on devait construire.

Le théorème démontré autorise une représentation matricielle.

PROPOSITION 4. *Si A est une matrice symétrique, il existe une matrice orthogonale S telle que* $S^{-1}AS$ *soit une matrice diagonale.*

---

*) Dans le « pire » des cas, toutes les racines du polynôme caractéristique coïncident, de sorte que $\mathbf{A}$ a une seule valeur propre.

En effet, la matrice A rapportée à une base orthonormée définit la transformation symétrique. Pour S on peut prendre la matrice de passage de cette base à la base construite dans le théorème. Rappelons que la matrice d'une transformation est diagonale par rapport à la base des vecteurs propres de cette transformation (proposition 6, § 4, ch. VI).

On a déjà étudié dans le théorème 1 du § 3, ch. IV, la transformation affine du plan, consistant en contraction (traction) suivant deux directions perpendiculaires. La généralisation d'une telle transformation dans l'espace euclidien n-dimensionnel des vecteurs est la contraction (traction) suivant n directions deux à deux perpendiculaires. Choisissons une base orthonormée de manière que ses vecteurs possèdent ces directions. Chaque vecteur de base e_i se transforme alors en vecteur $\lambda_i e_i$ qui lui est proportionnel, où λ_i est le coefficient de contraction. (Cette propriété représente justement la définition correcte de la contraction suivant des directions deux à deux perpendiculaires.) Rapportée à cette base, la matrice de la transformation étudiée est une matrice diagonale, et ses éléments diagonaux sont les coefficients de contraction. Vu que la matrice diagonale est symétrique et la base est orthonormée, la contraction suivant n directions deux à deux perpendiculaires est une transformation symétrique.

Inversement, en vertu du théorème 4, toute transformation symétrique dont les valeurs propres sont strictement positives est une contraction suivant n directions deux à deux perpendiculaires. A la valeur propre nulle, correspond non pas une contraction mais un projecteur orthogonal, tandis qu'à une valeur propre strictement négative est associé le produit d'une contraction par une symétrie.

Si λ^* est une valeur propre de multiplicité s, il lui correspond un sous-espace invariant $\mathscr{E}_s$ de dimension s. En effet, s'il n'en était pas ainsi, il n'existerait pas de base formée des vecteurs propres : la somme de toutes les multiplicités vaut n, quant au nombre de vecteurs linéairement indépendants associés à chaque valeur propre il ne peut dépasser sa multiplicité.

Pour $\lambda^* > 0$, la restriction de la transformation **A** à cet espace invariant $\mathscr{E}_s$ est une homothétie, autrement dit une contraction uniforme suivant toutes les directions, de coefficient λ^*.

Etudions maintenant la méthode de recherche pratique de la base dont l'existence est démontrée dans le théorème. Choisissons une base (de préférence, orthonormée) et considérons la matrice de cette transformation. Cherchons les racines de son polynôme caractéristique $\det(A - \lambda E) = 0$ et, pour chaque racine, déterminons les vecteurs propres en résolvant le système d'équations $(A - \lambda E)\boldsymbol{\xi} = \boldsymbol{o}$. Pour les racines simples, il reste à normer les solutions non triviales. Pour la racine de multiplicité s, on obtient un système fondamental de s solutions. Ce sont des vecteurs propres linéairement indépendants et en général non orthogonaux. On doit les orthogonaliser et normer.

3. Isomorphisme d'espaces euclidiens. DÉFINITION. Les espaces euclidiens $\mathscr{E}$ et $\mathscr{E}'$ sont dits *isomorphes* s'il existe une application linéaire bijective $\mathbf{A}: \mathscr{E} \to \mathscr{E}'$ telle que

$$(\mathbf{A}(x), \mathbf{A}(y)) = (x, y) \tag{8}$$

pour tous x et y de $\mathscr{E}$. L'application $\mathbf{A}$ est appelée *isomorphisme* d'espaces euclidiens.

Ainsi, le terme «isomorphisme» prend des significations variées suivant le contexte. S'il s'agit d'espaces euclidiens, l'isomorphisme, tout en conservant les résultats d'opérations linéaires, conserve aussi le produit scalaire.

Pour que deux espaces euclidiens soient isomorphes, il est évidemment nécessaire qu'ils soient de même dimension, sinon, ils ne sont pas isomorphes même en tant qu'espaces vectoriels. Il s'avère que cette condition est également suffisante.

THÉORÈME 5. *Deux espaces euclidiens de même dimension sont toujours isomorphes. Les espaces euclidiens de dimensions différentes ne sont pas isomorphes.*

Pour démontrer la première assertion, choisissons dans chacun des espaces considérés $\mathscr{E}$ et $\mathscr{E}'$ une base orthonormée. Définissons l'application $\mathbf{A}: \mathscr{E} \to \mathscr{E}'$ en faisant correspondre des vecteurs qui possèdent les mêmes colonnes de coordonnées dans les bases choisies. La matrice de cette application est la matrice unité, de sorte que l'application $\mathbf{A}$ est un isomorphisme des espaces $\mathscr{E}$ et $\mathscr{E}'$ considérés comme des espaces vectoriels. La formule (11) du § 1 entraîne que cette application conserve le produit scalaire.

Il est opportun de signaler que la condition (8) est très forte. On en déduit que $\mathbf{A}$ est une application linéaire et de plus injective. En effet, considérons un vecteur quelconque x de $\mathscr{E}$ et un nombre arbitraire α. Le carré scalaire du vecteur $\mathbf{A}(\alpha x) - \alpha\mathbf{A}(x)$ de $\mathscr{E}'$ peut être écrit sous la forme

$$(\mathbf{A}(\alpha x), \mathbf{A}(\alpha x)) - 2\alpha(\mathbf{A}(\alpha x), \mathbf{A}(x)) + \alpha^2(\mathbf{A}(x), \mathbf{A}(x)).$$

Compte tenu de (8), on voit que cette expression est égale à $(\alpha x, \alpha x) - 2\alpha(\alpha x, x) + \alpha^2(x, x)$, c'est-à-dire à zéro. Ainsi, $\mathbf{A}(\alpha x) = \alpha\mathbf{A}(x)$. De façon analogue, on démontre que $\mathbf{A}(x + y) = \mathbf{A}(x) + \mathbf{A}(y)$.

Supposons ensuite que le vecteur x appartient au noyau de l'applicaton $\mathbf{A}$, c'est-à-dire que $\mathbf{A}(x) = o$. Cela signifie que $(\mathbf{A}(x), \mathbf{A}(x)) = 0$ et, en vertu de (8), que $(x, x) = 0$. Donc, le noyau de $\mathbf{A}$ est le sous-espace nul, et $\mathbf{A}$ une injection.

Dans le cas général, l'application $\mathbf{A}$ satisfaisant à la condition (8) n'est pas bijective : elle peut être un isomorphisme de $\mathscr{E}$ sur un sous-espace de $\mathscr{E}'$.

Supposons que la dimension m de l'espace $\mathscr{E}'$ est égale à la dimension n de l'espace $\mathscr{E}$. **A** étant une injection, son rang r est égal à n et par suite, $m = n = r$. Selon la proposition 6 du § 3, ch. VI, **A** est un isomorphisme. On a ainsi démontré la proposition suivante.

PROPOSITION 5. *Une application de l'espace euclidien $\mathscr{E}_n$ dans l'espace euclidien $\mathscr{E}_n'$ de même dimension est un isomorphisme si elle conserve le produit scalaire.*

4. Transformation orthogonale. La transformation **A** de l'espace euclidien $\mathscr{E}$ est dite *orthogonale* si elle conserve le produit scalaire; c'est-à-dire si la condition (8) est satisfaite pour tous les vecteurs de $\mathscr{E}$.

Il ressort de la proposition 5 que la transformation orthogonale est un isomorphisme de $\mathscr{E}$ sur lui-même.

PROPOSITION 6. *La transformation adjointe d'une transformation orthogonale est égale à la transformation réciproque :* $\mathbf{A}^* = \mathbf{A}^{-1}$.

En effet, d'après la formule (8) on a $(x, \mathbf{A}^* \circ \mathbf{A}(y)) = (x, y)$ ou $(x, \mathbf{A}^* \circ \mathbf{A}(y) - y) = 0$. Cela signifie que le vecteur $\mathbf{A}^* \circ \mathbf{A}(y) - y$ est orthogonal à tout vecteur de l'espace et, par suite, est nul. L'égalité $\mathbf{A}^* \circ \mathbf{A}(y) = y$ étant vérifiée pour tous les vecteurs, la transformation $\mathbf{A}^* \circ \mathbf{A}$ estidentique, d'où la proposition 6.

COROLLAIRE. *Dans une base orthonormée, la matrice de la transformation orthogonale est orthogonale :* ${}^tAA = E$.

PROPOSITION 7. *Les racines du polynôme caractéristique d'une transformation orthogonale* **A** *(y compris les racines complexes) sont en valeur absolue égales à l'unité.*

Soient A la matrice de **A** dans une base orthonormée, et λ une racine (peut-être complexe) de l'équation $\det(A - \lambda E) = 0$. Le système d'équations $(A - \lambda E)\xi = o$, a, en général, une solution non triviale complexe. Il ressort de l'égalité $A\xi = \lambda\xi$ que ${}^t\bar{\xi}\,{}^t\bar{A} = \bar{\lambda}\,{}^t\bar{\xi}$. Multiplions à gauche chaque membre de la première égalité par le membre correspondant de la seconde : ${}^t\bar{\xi}\,{}^tA\,A\xi = \lambda\bar{\lambda}\,{}^t\bar{\xi}\xi$. Vu que $A = \bar{A}$, ${}^tAA = E$ et ${}^t\bar{\xi}\xi \neq 0$ on a $\lambda\bar{\lambda} = 1$, ce qu'il fallait démontrer.

PROPOSITION 8. *Si $\mathscr{E}'$ est un sous-espace invariant par la transformation orthogonale* **A**, *le supplémentaire orthogonal $(\mathscr{E}')^\perp$ du sous-espace $\mathscr{E}'$ est également invariant.*

Pour le démontrer, signalons que la restriction $\mathbf{A}'$ de la transformation **A** à $\mathscr{E}'$ est une transformation orthogonale et, partant, bijective. Elle est, en particulier, une surjection. Ainsi donc, si x appartient à $\mathscr{E}'$, il en est de même du vecteur $\mathbf{A}^{-1}(x)$.

Soit y un vecteur quelconque de $(\mathscr{E}')^\perp$. Il nous faut démontrer que son image $\mathbf{A}(y)$ est orthogonal à tout vecteur x de $\mathscr{E}'$. On a

$$(x, \mathbf{A}(y)) = (\mathbf{A}^{-1}(x), y) = 0,$$

puisque y est orthogonal à tout vecteur de $\mathscr{E}'$. La proposition est démontrée.

THÉORÈME 6. *Soit* **A** *une transformation orthogonale dans un espace euclidien $\mathscr{E}_n$ de dimension n. $\mathscr{E}_n$ est alors une somme directe d'espaces uni- et bidimensionnels invariants par* **A**.

DÉMONSTRATION. Toute transformation orthogonale dans un espace euclidien a au moins un espace invariant uni- ou bidimensionnel, vu que son polynôme caractéristique possède au moins une racine réelle ou complexe.

En nous appuyant sur ce fait, démontrons le théorème par récurrence. Pour les espaces uni- et bidimensionnels, l'assertion ne fait aucun doute. Supposons qu'on a déjà démontré le théorème pour les espaces de dimensions $k - 1$ et $k - 2$ et démontrons-le pour un espace de dimension k. Il existe dans $\mathscr{E}_k$ un sous-espace invariant uni- ou bidimensionnel $\mathscr{E}'$ et son supplémentaire orthogonal $(\mathscr{E}')^\perp$ est un sous-espace invariant de dimension $k - 1$ ou $k - 2$ respectivement. Appliquons l'hypothèse de récurrence à la restriction $\mathbf{A}'$ de la transformation $\mathbf{A}$ sur $(\mathscr{E}')^\perp$. Les sous-espaces $\mathscr{E}''$, $\mathscr{E}'''$, ..., en lesquels se décompose $(\mathscr{E}')^\perp$ seront aussi invariants par $\mathbf{A}$. Vu que $\mathscr{E}_k$ est la somme directe de $\mathscr{E}'$ et $(\mathscr{E}')^\perp$, et que $(\mathscr{E}')^\perp$ est la somme directe de $\mathscr{E}''$, $\mathscr{E}'''$, ..., l'espace $\mathscr{E}_k$ est la somme directe de $\mathscr{E}'$, $\mathscr{E}''$, $\mathscr{E}'''$, ... Le théorème est démontré.

On peut supposer que les sous-espaces invariants bidimensionnels ne renferment pas des sous-espaces invariants unidimensionnels. En effet, si un sous-espace invariant bidimensionnel $\mathscr{E}_2$ contient un sous-espace unidimensionnel $\mathscr{E}_1$, il contient aussi un second sous-espace invariant unidimensionnel $\mathscr{E}_1^\perp$ qui est le supplémentaire orthogonal du premier jusqu'à $\mathscr{E}_2$. Donc, $\mathscr{E}_2$ est la somme directe de $\mathscr{E}_1$ et $\mathscr{E}_1^\perp$ et dans la décomposition de $\mathscr{E}_n$ on peut remplacer $\mathscr{E}_2$ par deux termes $\mathscr{E}_1$ et $\mathscr{E}_1^\perp$.

Choisissons une base orthonormée dans chacun des sous-espaces invariants bi- et unidimensionnels en lesquels se décompose $\mathscr{E}_n$ et réunissons ces bases. On obtient la base orthonormée $\| \mathbf{e}_1, \ldots, \mathbf{e}_n \|$ dans $\mathscr{E}_n$. Construisons la matrice A de la transformation $\mathbf{A}$ par rapport à cette base.

Supposons que le vecteur de base $\mathbf{e}_i$ appartient à un espace invariant unidimensionnel, autrement dit est propre. En vertu de la proposition 5, la valeur propre associée est 1 ou -1, de sorte que tous les éléments de la i-ième colonne de la matrice sont nuls, à l'exception de l'élément α_i^i sur la diagonale principale, égal à 1 ou -1.

Considérons les vecteurs de base $\mathbf{e}_k$ et $\mathbf{e}_{k+1}$ qui appartiennent au sous-espace invariant bidimensionnel $\mathscr{E}'$. Les vecteurs $\mathbf{A}(\mathbf{e}_k)$ et $\mathbf{A}(\mathbf{e}_{k+1})$ ne se décomposent que suivant les vecteurs $\mathbf{e}_k$ et $\mathbf{e}_{k+1}$, et par suite, tous les éléments dans la k-ième et la $(k + 1)$-ième colonne de la matrice A sont nuls à l'exception du bloc d'ordre deux

$$\left\| \begin{matrix} \alpha_k^k & \alpha_{k+1}^k \\ \alpha_k^{k+1} & \alpha_{k+1}^{k+1} \end{matrix} \right\|$$

de la diagonale principale. Ce bloc est la matrice de la transformation $\mathbf{A}'$ qui est la restriction de $\mathbf{A}$ au sous-espace $\mathscr{E}'$. Selon la proposition 6 du § 2, ch. IV, la matrice de la transformation orthogonale dans un espace bidimensionnel prend par rapport à la base orthonormée la forme

$$\left\| \begin{matrix} \cos\varphi & \mp\sin\varphi \\ \sin\varphi & \pm\cos\varphi \end{matrix} \right\|,$$

où les signes supérieurs sont pris pour les transformations orthogonales de première espèce et les signes inférieurs pour les transformations de seconde espèce. Si l'on prend les signes inférieurs, la matrice devient symétrique, de sorte que la transformation orthogonale de seconde espèce est symétrique, par suite, possède deux sous-espaces invariants unidimensionnels. Le résultat obtenu contredit l'hypothèse faite plus haut. Donc, il existe un nombre φ_k tel que

$$\left\| \begin{matrix} \alpha_k^k & \alpha_{k+1}^k \\ \alpha_k^{k+1} & \alpha_{k+1}^{k+1} \end{matrix} \right\| = \left\| \begin{matrix} \cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & \cos\varphi_k \end{matrix} \right\|.$$

Laissons au lecteur le soin d'écrire la forme générale de la matrice A.

§ 3. Notion d'espace unitaire

1. Définition. On montrera dans ce paragraphe comment se définit le produit scalaire dans des espaces vectoriels complexes. On s'abstiendra cependant de fournir des démonstrations qui peuvent être obtenues, avec de légères modifications, à partir des démonstrations données pour les propositions correspondantes au § 1. Tous les nombres sont ici en général des nombres complexes.

Considérons un espace vectoriel complexe $\mathscr{L}$ et supposons qu'il existe une relation qui fait correspondre à chaque couple de vecteurs x et y le nombre (x, y). Il s'avère que les axiomes mentionnés dans la définiton de l'espace euclidien ne peuvent pas être vérifiés. En effet, soit x un vecteur non nul. L'espace étudié est muni de la multiplication d'un vecteur par un nombre complexe, de sorte qu'on peut prendre le vecteur ix, où i est l'unité imaginaire. Si les axiomes 1° et 2° sont vérifiés, on a l'égalité

$$(ix, ix) = -(x, x).$$

Avec un produit positif à droite, le produit à gauche est négatif, ce qui contredit l'axiome 4°.

Ce fait oblige d'introduire d'autres définitions du produit scalaire dans des espaces complexes. Dans l'une d'elles, on remplace l'axiome 4° par une condition moins rigoureuse : on exige que $(x, y) = 0$ quel que soit x implique $y = 0$. L'espace vectoriel complexe muni d'un produit scalaire ainsi défini s'appelle *espace euclidien complexe*. Les espaces euclidiens complexes sont relativement peu utilisés. Beaucoup plus souvent dans les applications, on rencontre des espaces dits unitaires.

DÉFINITION. L'espace vectoriel complexe $\mathscr{L}$ est dit *unitaire* (ou *hermitien*) s'il existe une relation qui fait correspondre à tout couple de vecteurs x et y de $\mathscr{L}$ un nombre complexe (x, y), et telle que sont vérifiés les axiomes suivants, quels que soient les vecteurs x, y et z et le nombre α :

1° $(x, y) = \overline{(y, x)}$, c'est-à-dire que lorsqu'on permute les facteurs, on remplace le produit scalaire par le nombre complexe conjugué;

2° $(\alpha x, y) = \alpha(x, y)$;

3° $(x + y, z) = (x, z) + (y, z)$;

4° $(x, x) > 0$, si $x \neq o$.

Le nombre (x, y) est appelé *produit scalaire* des vecteurs x et y.

Remarquons que pour tout vecteur x on a $(x, x) = \overline{(x, x)}$ et, par suite, le carré scalaire du vecteur est toujours un nombre réel. L'axiome 4° exige que ce nombre soit strictement positif pour $x \neq o$.

Les axiomes 1° et 2° impliquent la règle suivante de mise en facteur scalaire :

$$(x, \alpha y) = \overline{(\alpha y, x)} = \overline{\alpha (y, x)} = \bar{\alpha}\overline{(y, x)}$$

(comp. (4), § 7, ch. V). En définitive,

$$(x, \alpha y) = \bar{\alpha}(x, y). \tag{1}$$

La longueur du vecteur et l'angle de deux vecteurs se définissent par les mêmes formules que dans le cas réel. La longueur du vecteur est toujours définie par le nombre réel positif. L'angle est en général défini par le nombre complexe.

Les vecteurs sont dits *orthogonaux* si leur produit scalaire est nul. Seul le vecteur nul est orthogonal à tout vecteur.

Signalons qu'on a l'inégalité

$$(x, x)(y, y) \geqslant |(x, y)|^2 = (x, y)(y, x).$$

Elle se démontre de la même façon que l'inégalité (7) du § 1, compte tenu de ce que

$$(\alpha x + \beta y, \alpha x + \beta y) = \alpha\bar{\alpha}(x, x) + \alpha\bar{\beta}(x, y) + \beta\bar{\alpha}(y, x) + \beta\bar{\beta}(y, y).$$

EXEMPLE 1. L'espace vectoriel complexe des matrices-colonnes à n éléments devient un espace unitaire de dimension n si le produit scalaire des matrices-colonnes $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ est défini par la formule

$$(\boldsymbol{\xi}, \boldsymbol{\eta}) = \xi^1\bar{\eta}^1 + \ldots + \xi^n\bar{\eta}^n. \tag{2}$$

En effet, en appliquant cette formule, on a aussi

$$(\boldsymbol{\eta}, \boldsymbol{\xi}) = \eta^1\bar{\xi}^1 + \ldots + \eta^n\bar{\xi}^n.$$

Il ressort des formules (3) et (4) du § 7, ch. V, que $(\boldsymbol{\xi}, \boldsymbol{\eta}) = \overline{(\boldsymbol{\eta}, \boldsymbol{\xi})}$.

Les axiomes 2° et 3° découlent des propriétés de la multiplication des matrices si l'on remarque que le second membre de (2) est le produit ${}^t\boldsymbol{\xi}\bar{\boldsymbol{\eta}}$, où $\bar{\boldsymbol{\eta}}$ est la matrice-colonne composée des éléments $\bar{\eta}^1, \ldots, \bar{\eta}^n$. Enfin,

$$(\boldsymbol{\xi}, \boldsymbol{\xi}) = \xi^1\overline{\xi^1} + \ldots + \xi^n\overline{\xi^n} = |\xi^1|^2 + \ldots + |\xi^n|^2 \tag{3}$$

et, par suite, le carré scalaire de la matrice-colonne est positif; il est égal à zéro si la matrice-colonne est nulle.

EXEMPLE 2. On peut construire un espace unitaire unidimensionnel de la façon suivante. Considérons l'ensemble de tous les vecteurs d'un plan, muni de l'opération d'addition définie par la règle du parallélogramme.

Pour définir l'opération de multiplication par un nombre complexe, choisissons une base orthonormée (dans le sens habituel) $\|\mathbf{e}_1, \mathbf{e}_2\|$. On appellera produit du vecteur $\mathbf{x}$ de composantes ξ^1, ξ^2 par le nombre $\alpha + i\beta$ le vecteur de composantes $\alpha\xi^1 - \beta\xi^2$ et $\alpha\xi^2 + \beta\xi^1$. Le sens de cette définition est le suivant. A chaque vecteur on peut faire correspondre un nombre $\xi^1 + i\xi^2$; ceci étant, la correspondance entre les nombres et les vecteurs est biunivoque : à tout vecteur est associé un nombre et un seul et à tout nombre est associé un vecteur et un seul. Au produit $(\alpha + i\beta)\mathbf{x}$ est associé le

nombre $(\alpha + i\beta)(\xi^1 + i\xi^2)$. Signalons qu'à la somme des vecteurs correspond la somme des nombres associés à ces vecteurs.

Voyons si les axiomes de l'espace vectoriel y sont vérifiés. Les quatre premiers axiomes se rapportant à l'addition des vecteurs sont évidemment vérifiés. Les égalités $(\lambda + \mu)\mathbf{x} = \lambda\mathbf{x} + \mu\mathbf{x}$ et $\lambda(\mathbf{x} + \mathbf{y}) = \lambda\mathbf{x} + \lambda\mathbf{y}$ se déduisent de la distributivité de la multiplication des nombres complexes. L'égalité $\lambda(\mu\mathbf{x}) = (\lambda\mu)\mathbf{x}$ découle de l'associativité de la multiplication. Il est aussi évident que $1\mathbf{x} = \mathbf{x}$. On a donc un espace vectoriel complexe. Sa dimension est 1, car tout vecteur $\mathbf{x}$ est égal à $(\xi^1 + i\xi^2)\mathbf{e}_1$, où $\xi^1 + i\xi^2$ est un nombre complexe défini par le vecteur $\mathbf{x}$. La base est représentée par le vecteur $\mathbf{e}_1$.

Définissons le produit scalaire des vecteurs $\mathbf{x} = \lambda\mathbf{e}_1$ et $\mathbf{y} = \mu\mathbf{e}_1$ par la formule $(\mathbf{x}, \mathbf{y}) = \lambda\bar{\mu}$. Il est aisé de vérifier que tous les axiomes de la multiplication scalaire pour un espace unitaire sont satisfaits.

La longueur unitaire du vecteur $(1 + i)\mathbf{e}_1$ est $\sqrt{2}$. Le produit scalaire $(\mathbf{e}_1, \mathbf{e}_2) = (\mathbf{e}_1, i\mathbf{e}_1) = -i$, bien que relativement au produit scalaire dans le plan réel les vecteurs $\mathbf{e}_1$ et $\mathbf{e}_2$ sont perpendiculaires.

2. Propriétés des espaces unitaires. Toutes les propriétés étudiées des espaces euclidiens se rapportent avec d'infimes modifications aux espaces unitaires.

Il existe dans un espace unitaire de dimension finie une base orthonormée, c'est-à-dire une base formée de vecteurs de longueur unité deux à deux orthogonaux. Une telle base peut être obtenue à partir d'une base quelconque par orthogonalisation.

Le produit scalaire s'exprime en fonction des composantes des facteurs par rapport à une base orthonormée suivant la formule

$$(x, y) = \xi^1\overline{\eta^1} + \ldots + \xi^n\overline{\eta^n}.$$

Pour une base arbitraire on introduit la matrice de Gram Γ dont les éléments sont les produits scalaires des vecteurs de base pris deux à deux. Le produit scalaire des vecteurs x et y dont les colonnes de coordonnées sont $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ s'exprime par la formule

$$(x, y) = {}^t\boldsymbol{\xi}\Gamma\bar{\boldsymbol{\eta}}.$$

Vu que $(e_i, e_j) = \overline{(e_j, e_i)}$ dans un espace unitaire, la matrice de Gram satisfait à la condition

$$^t\Gamma = \bar{\Gamma}. \tag{4}$$

(Ici et plus loin le trait au-dessus de la matrice signifie le passage de tous ses éléments aux nombres complexes conjugués.)

DÉFINITION. Toute matrice satisfaisant à la condition (4) est dite *hermitienne*.

La matrice de passage S d'une base orthonormée de l'espace unitaire à

une autre base orthonormée doit vérifier l'égalité

$$ {}^{t}S\bar{S} = E. \tag{5}$$

Cela signifie que $S^{-1} = {}^{t}\bar{S}$, d'où il vient que

$$\bar{S}\,{}^{t}S = E.$$

Définition. La matrice S est dite *unitaire* si elle vérifie l'égalité (5).

Remarquons que l'égalité (5) et la formule (7) du § 7, ch. V, entraînent

$$\det({}^{t}S\bar{S}) = \det {}^{t}S \det \bar{S} = (\det S)(\overline{\det S}) = |\det S|^2 = 1,$$

de sorte que le déterminant de la matrice unitaire est un nombre complexe de module égal à l'unité.

Le supplémentaire orthogonal d'un sous-espace de l'espace unitaire se définit de la même façon que dans l'espace euclidien. De la même manière on démontre que le supplémentaire orthogonal du sous-espace de dimension k est un sous-espace de dimension $n - k$.

3. Transformations auto-adjointes et unitaires. La transformation d'un espace unitaire est dite *auto-adjointe* si pour tous vecteurs x et y est vérifiée l'égalité

$$(\mathbf{A}(x), y) = (x, \mathbf{A}(y)).$$

Il découle de cette définition que la transformation de l'espace unitaire est auto-adjointe si et seulement si sa matrice est hermitienne dans toute base orthonormée. Aux transformations auto-adjointes d'espaces unitaires se rapportent sans changements les théorèmes 1 à 4 du § 2.

On dit qu'une transformation de l'espace unitaire est *unitaire* si elle satisfait pour tous vecteurs x et y à la condition

$$(\mathbf{A}(x), \mathbf{A}(y)) = (x, y).$$

On vérifie aisément que la transformation est unitaire si et seulement si sa matrice est unitaire dans toute base orthonormée.

CHAPITRE VIII

FONCTIONS SUR L'ESPACE VECTORIEL

§ 1. Fonctions linéaires

1. Définition d'une fonction. De même qu'au chapitre VI, on considère un espace vectoriel quelconque. Dans le cas où la différence entre les espaces réels et complexes serait essentielle, on introduira des données supplémentaires. Si le terme « nombre » n'est pas précisé, on sous-entend un nombre complexe dans le cas de l'espace complexe, et un nombre réel dans le cas de l'espace réel. En règle générale, le produit scalaire n'est pas introduit, mais de nombreux résultats se rapportent aux espaces euclidiens.

DÉFINITION. On dira qu'une *fonction* (d'un vecteur) est définie sur l'espace vectoriel $\mathscr{L}$ si à chaque vecteur x de $\mathscr{L}$ est associé un nombre. Une *fonction de deux vecteurs* est définie sur $\mathscr{L}$ si à chaque couple ordonné de vecteurs x, y de $\mathscr{L}$ est associé un nombre.

Le nombre que la fonction $\mathfrak{f}$ fait correspondre au vecteur x s'appelle *valeur de la fonction* $\mathfrak{f}$ *sur* x et se note $\mathfrak{f}(x)$. De façon analogue on définit la valeur $\mathbf{g}(x, y)$ de la fonction $\mathbf{g}$ de deux vecteurs.

Les fonctions sur les espaces de dimension infinie sont généralement appelées *fonctionnelles*.

Soit $\mathscr{L}_n$ un espace de dimension n rapporté à une base. A chaque vecteur x de $\mathscr{L}_n$ on peut alors associer ses n composantes $\xi^1, \ldots, \xi^n$. Rappelons qu'en analyse mathématique on appelle *fonction de n variables* une relation faisant correspondre un nombre à tout n-uple ordonné de nombres $(\xi^1, \ldots, \xi^n)$ de l'ensemble donné. Ainsi donc, la base étant choisie, la fonction $\mathfrak{f}$ sur l'espace vectoriel $\mathscr{L}_n$ se détermine par la fonction de n variables définie pour tous les n-uples possibles $(\xi^1, \ldots, \xi^n)$. Si on change de base, on associe à tout vecteur x ses nouvelles composantes $\xi'^1, \ldots, \xi'^n$ et, par suite, la fonction $\mathfrak{f}$ se détermine par une nouvelle fonction de n variables.

2. Fonctions linéaires. DÉFINITION. La fonction $\mathfrak{f}$ sur un espace vectoriel $\mathscr{L}$ est dite *linéaire* si pour tous vecteurs x et y de $\mathscr{L}$ et tout nombre α on a

$$\mathfrak{f}(x + y) = \mathfrak{f}(x) + \mathfrak{f}(y), \qquad \mathfrak{f}(\alpha x) = \alpha \mathfrak{f}(x). \tag{1}$$

Le lecteur notera que la notion de fonction linéaire sur un espace vectoriel ne lui est pas étrangère. C'est exactement la même chose qu'une appli-

cation linéaire de l'espace vectoriel donné dans un espace arithmétique unidimensionnel (comp. exemple 3), § 3, ch. VI).

EXEMPLE 1. La fonction qui associe à chaque vecteur le nombre zéro est linéaire. Une fonction faisant correspondre à tous les vecteurs un même nombre non nul ne peut être linéaire car pour chaque fonction linéaire on a $\mathbf{f}(o) = 0$. On invite le lecteur à vérifier ces assertions.

EXEMPLE 2. Soit $\mathscr{E}_n$ un espace euclidien de dimension n et soit a un vecteur fixé de $\mathscr{E}_n$. A chaque vecteur x de $\mathscr{E}_n$ on peut alors faire correspondre le nombre $\zeta = (a, x)$. Les égalités (1) sont évidemment vérifiées et on a donc une fonction linéaire sur $\mathscr{E}_n$.

EXEMPLE 3. Soit dans l'espace vectoriel $\mathscr{L}_n$ une base $\|e_1, \ldots, e_n\|$. Associons à chaque vecteur x sa i-ième composante ξ^i dans cette base. Cette correspondance est évidemment une fonction linéaire sur $\mathscr{L}_n$. On la notera $\mathbf{p}^i$. On peut ainsi construire n fonctions $\mathbf{p}^1, \ldots, \mathbf{p}^n$. Elles dépendent évidemment de la base choisie.

EXEMPLE 4. Considérons un espace vectoriel $\mathscr{L}$ de dimension infinie, constitué de fonctions d'une seule variable indépendante ξ, définies et continues pour $0 \leqslant \xi \leqslant 1$. Soit $v(\xi)$ une fonction fixée de $\mathscr{L}$, par exemple $v(\xi) = \sin \xi$. On peut alors faire correspondre à chaque fonction $u(\xi)$ de $\mathscr{L}$ le nombre

$$\zeta = \int_0^1 v(\xi)u(\xi)d\xi.$$

Il est évident que cette correspondance est une fonctionnelle linéaire. D'ailleurs, si l'on se rappelle l'exemple 3 du § 1, ch. VII, il devient évident que cette fonctionnelle est construite de la même façon que la fonction linéaire de l'exemple 2.

On obtiendra encore une fonctionnelle linéaire sur le même espace $\mathscr{L}$ si l'on associe à chaque fonction $u(\xi)$ de $\mathscr{L}$ sa valeur $u(0)$ pour $\xi = 0$.

Soit un espace vectoriel quelconque $\mathscr{L}_n$ de dimension n rapporté à une base $\|e_1, \ldots, e_n\|$. La valeur de la fonction $\mathbf{f}$ sur le vecteur x de $\mathscr{L}_n$ peut être écrite au moyen des composantes de ce vecteur $\xi^1, \ldots, \xi^n$:

$$\mathbf{f}(x) = \mathbf{f}(\xi^1 e_1 + \ldots + \xi^n e_n) = \xi^1 \mathbf{f}(e_1) + \ldots + \xi^n \mathbf{f}(e_n).$$

Les nombres $\mathbf{f}(e_1), \ldots, \mathbf{f}(e_n)$ sont indépendants du vecteur x et ne sont définis que par la fonction $\mathbf{f}$ et la base $\|e_1, \ldots, e_n\|$. On a ainsi démontré la proposition suivante.

PROPOSITION 1. *Toute fonction linéaire sur un espace vectoriel de dimension n rapporté à une base* $\|e_1, \ldots, e_n\|$ *se définit par un polynôme linéaire*

homogène

$$\mathbf{f}(x) = \varkappa_1 \xi^1 + \ldots + \varkappa_n \xi^n, \tag{2}$$

où $\xi^1, \ldots, \xi^n$ *sont les composantes du vecteur x par rapport à cette base, et* $\varkappa_1, \ldots, \varkappa_n$ *les coefficients du polynôme qui sont égaux aux valeurs de la fonction sur les vecteurs de base.*

Il est commode de dire que les valeurs de la fonction $\mathbf{f}$ sur les vecteurs de la base $\boldsymbol{e}$ sont les *composantes* (ou *coefficients*) de la fonction $\mathbf{f}$ dans la base $\boldsymbol{e}$. La matrice de l'application linéaire d'un espace de dimension n dans un espace unidimensionnel est de type $(1, n)$, autrement dit est une matrice-ligne à n éléments. Dans le cas considéré, c'est la matrice-ligne $\|\varkappa_1 \ldots \varkappa_n\|$. Laissons au soin du lecteur de le vérifier. La formule (2) s'écrit sous forme matricielle ainsi :

$$\mathbf{f}(x) = \|\varkappa_1 \ldots \varkappa_n\| \begin{Vmatrix} \xi^1 \\ \ldots \\ \xi^n \end{Vmatrix} = \boldsymbol{\varkappa\xi}. \tag{3}$$

On s'aperçoit aisément que chaque matrice-ligne $\boldsymbol{\varkappa}$ définit par la formule (3) une fonction linéaire. En effet, $\boldsymbol{\varkappa}(\boldsymbol{\xi} + \boldsymbol{\eta}) = \boldsymbol{\varkappa\xi} + \boldsymbol{\varkappa\eta}$ et $\boldsymbol{\varkappa}(\alpha\boldsymbol{\xi}) = \alpha(\boldsymbol{\varkappa\xi})$.

La formule (6) du § 3, ch. VI, exprime la matrice de l'application par rapport aux nouvelles bases au moyen de l'ancienne matrice de cette application et les matrices de passage aux nouvelles bases. Etant donné que dans l'espace arithmétique unidimensionnel la base est fixée une fois pour toutes, cette formule prend pour une fonction linéaire la forme

$$\boldsymbol{\varkappa}' = \boldsymbol{\varkappa} S. \tag{4}$$

$\boldsymbol{\varkappa}$ est ici la matrice-ligne des coefficients de la fonction dans la base $\boldsymbol{e}$, et $\boldsymbol{\varkappa}'$ la matrice-ligne de ses coefficients dans la base $\boldsymbol{e}' = \boldsymbol{e}S$. Il est évident que la formule (4) peut être obtenue directement. En effet, écrivons $\mathbf{f}(x)$ dans chacune des deux bases : $\mathbf{f}(x) = \boldsymbol{\varkappa\xi} = \boldsymbol{\varkappa}'\boldsymbol{\xi}'$. D'où, selon la formule (3) du § 1, ch. VI, $\boldsymbol{\varkappa} S\boldsymbol{\xi}' = \boldsymbol{\varkappa}'\boldsymbol{\xi}'$ ou $(\boldsymbol{\varkappa} S - \boldsymbol{\varkappa}')\boldsymbol{\xi}' = \boldsymbol{o}$, avec $\boldsymbol{\xi}'$ une matrice-colonne arbitraire. En portant successivement à sa place chaque colonne de la matrice unité, on s'apercevra que chaque élément de la matrice-ligne $\boldsymbol{\varkappa} S - \boldsymbol{\varkappa}'$ vaut zéro.

3. Espace dual. Dans le chapitre VI, on a introduit les définitions de la somme d'applications linéaires et du produit d'une application linéaire par un nombre. Rapportées aux fonctions linéaires ces définitions sont formulées ainsi :

DÉFINITION. On appelle *somme des fonctions linéaires* $\mathbf{f}$ et $\mathbf{g}$ la fonction $\mathbf{h}$ dont la valeur sur chaque vecteur x est définie par l'égalité $\mathbf{h}(x) = \mathbf{f}(x) + \mathbf{g}(x)$. On appelle *produit de la fonction linéaire* $\mathbf{f}$ *par un nombre* α

la fonction $\mathbf{g}$ dont la valeur sur chaque vecteur x est définie par l'égalité $\mathbf{g}(x) = \alpha\mathbf{f}(x)$.

PROPOSITION 2. *Soient* $\mathbf{f}$ *et* $\mathbf{g}$ *des fonctions linéaires sur l'espace vectoriel* $\mathscr{L}$, *et* $\boldsymbol{x}$ *et* $\boldsymbol{\lambda}$ *leurs matrices-lignes des coefficients dans une base* e. *La somme* $\mathbf{f} + \mathbf{g}$ *est alors une fonction linéaire dont la matrice-ligne des coefficients par rapport à la base* e *est* $\boldsymbol{x} + \boldsymbol{\lambda}$. *De même, pour un nombre quelconque* α, *le produit* $\alpha\mathbf{f}$ *est une fonction linéaire dont la matrice-ligne des coefficients par rapport à la base* e *est* $\alpha\boldsymbol{x}$.

Pour des applications linéaires quelconques on l'a démontré dans le point 6 du § 3, ch. VI. Reproduisons, toutefois, cette démonstration pour le cas de la somme de fonctions linéaires. Pour un vecteur arbitraire x, les valeurs de $\mathbf{f}$ et de $\mathbf{g}$ s'écrivent dans la base e sous forme de $\boldsymbol{x\xi}$ et $\boldsymbol{\lambda\xi}$. La valeur de la somme $\mathbf{f} + \mathbf{g}$ sur le même vecteur est alors égale à $\boldsymbol{x\xi} + \boldsymbol{\lambda\xi} = (\boldsymbol{x} + \boldsymbol{\lambda})\boldsymbol{\xi}$. Il s'ensuit que $\mathbf{f} + \mathbf{g}$ est une fonction linéaire dont la matrice-ligne des coefficients est $\boldsymbol{x} + \boldsymbol{\lambda}$.

PROPOSITION 3. *L'ensemble* $\mathscr{L}^*$ *de toutes les fonctions linéaires sur l'espace vectoriel* $\mathscr{L}_n$ *de dimension* n, *muni des opérations d'addition et de multiplication par un nombre introduites ci-dessus est un espace vectoriel de dimension* n.

En effet, il existe une application bijective de l'ensemble $\mathscr{L}^*$ sur l'ensemble des matrices-lignes à n éléments. Selon la proposition 2, l'image par cette application de la somme de fonctions est la somme des matrices-colonnes, et l'image du produit de la fonction par un nombre est le produit de la matrice-ligne par ce nombre. Etant donné que les axiomes de l'espace vectoriel sont vérifiés pour les opérations sur les matrices-lignes, ils le seront pour les opérations dans $\mathscr{L}^*$. Donc, $\mathscr{L}^*$ est un espace vectoriel isomorphe à l'espace des matrices-lignes à n éléments.

DÉFINITION. L'espace vectoriel $\mathscr{L}^*$ de toutes les fonctions linéaires définies sur l'espace vectoriel $\mathscr{L}$ est dit *dual* de l'espace $\mathscr{L}$.

Choisissons dans l'espace $\mathscr{L}_n$ une base e et considérons des fonctions linéaires $\mathbf{p}^i$ ($i = 1, \dots, n$) définies par les égalités $\mathbf{p}^i(x) = \xi^i$, où ξ^i est la i-ième composante du vecteur x (comp. exemple 3). Cela signifie que

$$\mathbf{p}^i(e_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases} \tag{5}$$

c'est-à-dire que la matrice-ligne des coefficients de la fonction $\mathbf{p}^i$ est la i-ième ligne de la matrice unité. Il en découle aussitôt que les fonctions $\mathbf{p}^1, \dots, \mathbf{p}^n$ sont linéairement indépendantes.

La matrice-ligne $\boldsymbol{x} = \|x_1 \dots x_n\|$ se décompose suivant les lignes de la matrice unité, avec coefficients $x_1, \dots, x_n$. Cela signifie que l'élément $\mathbf{f}$ de

l'espace $\mathscr{L}^*$ dont la matrice-ligne des coefficients est $\|x_1 \ldots x_n\|$ a pour décomposition

$$\mathbf{f} = x_1 \mathbf{p}^1 + \ldots + x_n \mathbf{p}^n. \tag{6}$$

Ainsi, $\|\mathbf{p}^1, \ldots, \mathbf{p}^n\|$ est la base dans l'espace $\mathscr{L}_n^*$.

DÉFINITION. La base $\|\mathbf{p}^1, \ldots, \mathbf{p}^n\|$ de l'espace $\mathscr{L}_n^*$ définie par la formule (5) est dite *duale* (ou *biorthogonale*) de la base $\|e_1, \ldots, e_n\|$ de l'espace $\mathscr{L}_n$.

Considérons la matrice-colonne $\boldsymbol{p}$ composée des fonctions $\mathbf{p}^i$. On peut maintenant présenter la décomposition (6) sous la forme matricielle :

$$\mathbf{f} = \|x_1 \ldots x_n\| \begin{Vmatrix} \mathbf{p}^1 \\ \ldots \\ \mathbf{p}^n \end{Vmatrix} = \boldsymbol{x}\boldsymbol{p}. \tag{7}$$

Si l'on convenait d'écrire les composantes d'un vecteur de l'espace $\mathscr{L}_n^*$ sous forme de matrice-colonne, la formule (7) se présenterait comme suit : $\mathbf{f} = {}^t\boldsymbol{p}\,{}^t\boldsymbol{x}$.

Supposons que les bases $\boldsymbol{e}$ et $\boldsymbol{e}'$ de l'espace $\mathscr{L}_n$ sont liées par l'égalité $\boldsymbol{e}' = \boldsymbol{e}S$. Cherchons la matrice de passage entre les bases duales $\boldsymbol{p}$ et $\boldsymbol{p}'$. A cet effet, écrivons la formule (4) sous la forme (3), § 1, ch. VI, en la résolvant relativement aux anciennes composantes et en écrivant les composantes sous forme de matrice-colonne. Il vient

$${}^t\boldsymbol{x} = {}^t(S^{-1})\,{}^t\boldsymbol{x}'.$$

On voit donc que la matrice de passage de la base $\boldsymbol{p}$ à la base $\boldsymbol{p}'$ dans l'espace $\mathscr{L}_n^*$ est la matrice ${}^t(S^{-1})$, c'est-à-dire qu'on a l'égalité ${}^t\boldsymbol{p}' = {}^t\boldsymbol{p}\,{}^t(S^{-1})$. Si l'on revient, pour l'espace $\mathscr{L}_n^*$, à l'écriture des éléments de base sous forme de matrice-colonne, la dépendance entre les bases prendra la forme

$$\boldsymbol{p} = S\boldsymbol{p}'. \tag{8}$$

L'espace $\mathscr{L}_n^*$ est un espace vectoriel comme tous les autres, si bien qu'il possède un espace dual $\mathscr{L}_n^{**}$ dont les éléments sont des fonctions linéaires sur $\mathscr{L}_n^*$.

PROPOSITION 4. *L'espace $\mathscr{L}^{**}$ peut être identifié à $\mathscr{L}_n$.*

DÉMONSTRATION. Fixons un vecteur x de $\mathscr{L}_n$ et faisons correspondre à chaque élément $\mathbf{f}$ de $\mathscr{L}_n^*$ le nombre $\mathbf{f}(x)$. On peut donc interpréter x comme une fonction sur $\mathscr{L}_n^*$. Cette fonction est linéaire. En effet, $(\mathbf{f} + \mathbf{g})(x) = \mathbf{f}(x) + \mathbf{g}(x)$ et, par suite, à la somme des éléments de $\mathscr{L}_n^*$ la fonction x fait correspondre la somme des nombres associés à ces éléments. D'une façon analogue, l'égalité $\alpha\mathbf{f}(x) = \alpha\mathbf{f}(x)$ veut dire qu'au produit de l'élément $\mathbf{f}$ par α la fonction x fait correspondre le produit de α par le nombre associé à $\mathbf{f}$. Ainsi donc, x peut être interprété comme un élément de $\mathscr{L}_n^{**}$.

Démontrons que l'espace $\mathscr{L}_n$ tout entier peut être identifié à un sous-

espace de $\mathscr{L}_n^{**}$. Il suffit pour cela de démontrer que la somme et le produit par un nombre des vecteurs de $\mathscr{L}_n$ se confondent avec la somme et le produit par un nombre de ces vecteurs interprétés comme fonctions linéaires sur $\mathscr{L}_n^*$. Or, ce fait est évident. Pour la somme par exemple, ceci est équivalent à la condition $\mathbf{f}(x + y) = \mathbf{f}(x) + \mathbf{f}(y)$ qui est satisfaite pour tous x, y de $\mathscr{L}_n$ et tout $\mathbf{f}$ de $\mathscr{L}_n^*$.

La coïncidence de $\mathscr{L}_n$ et de $\mathscr{L}_n^{**}$ découle maintenant de l'égalité de leurs dimensions, en vertu de la proposition 2 du § 2, ch. VI.

La proposition 4 signifie en fait qu'entre $\mathscr{L}_n$ et $\mathscr{L}_n^{**}$ il existe un isomorphisme canonique, indépendant du choix de la base.

4. Fonctions linéaires sur un espace euclidien. Le choix d'une base dans l'espace vectoriel $\mathscr{L}$ établit un isomorphisme entre $\mathscr{L}$ et $\mathscr{L}^*$. Si l'espace $\mathscr{L}$ est euclidien, l'isomorphisme entre $\mathscr{L}$ et son dual $\mathscr{L}^*$ peut être établi canoniquement sans tenir compte de la base. Dans l'exemple 2 on a montré précisément que la formule $\mathbf{f}(x) = (a, x)$ permet de faire correspondre à chaque vecteur a de l'espace euclidien une fonction linéaire $\mathbf{f}$ définie sur cet espace.

Appelons le vecteur a *vecteur associé* à la fonction $\mathbf{f}(x) = (a, x)$ et cherchons une relation entre la matrice-ligne $\boldsymbol{x}$ des coefficients de cette fonction et la colonne de coordonnées du vecteur a dans une base $\boldsymbol{e}$. On a $\mathbf{f}(x) =$ $= \boldsymbol{x}\boldsymbol{\xi} = {}^t\boldsymbol{\alpha}\Gamma\boldsymbol{\xi}$, où $\boldsymbol{\alpha}$ et $\boldsymbol{\xi}$ sont les colonnes de coordonnées des vecteurs a et x, et Γ est la matrice de Gram de la base $\boldsymbol{e}$. Les coefficients de la fonction linéaire étant définis univoquement (ce sont ses valeurs sur les vecteurs de base), la dernière égalité entraîne

$$\boldsymbol{x} = {}^t\boldsymbol{\alpha}\Gamma, \quad \text{ou} \quad {}^t\boldsymbol{x} = \Gamma\boldsymbol{\alpha}.$$

Cette dernière formule peut être interprétée comme expression analytique de l'application linéaire $\Gamma : \mathscr{E}_n \to \mathscr{E}_n^*$ par rapport aux bases $\boldsymbol{e}$ et $\boldsymbol{p}$, la base $\boldsymbol{p}$ étant la base duale de $\boldsymbol{e}$. Vu que $\det \Gamma \neq 0$, l'application Γ est un isomorphisme de $\mathscr{E}_n$ sur $\mathscr{E}_n^*$, les deux espaces étant considérés comme espaces vectoriels.

Il en découle, en particulier, que pour toute fonction linéaire $\mathbf{f}$ sur $\mathscr{E}_n$ il existe un vecteur associé f tel que $\mathbf{f}(x) = (f, x)$.

On n'a pas encore muni l'espace $\mathscr{E}_n^*$ de la multiplication scalaire. Mais on peut le faire à l'aide de la formule $(\Gamma(a), \Gamma(b)) = (a, b)$. Ceci étant, l'application Γ devient un isomorphisme entre espaces euclidiens.

Ainsi donc, il existe entre l'espace euclidien $\mathscr{E}_n$ et son dual un isomorphisme bien défini, lié au produit scalaire, qui permet d'identifier ces espaces. Cette identification est courante.

Considérons les vecteurs $p^i = \Gamma^{-1}(\mathbf{p}^i)$ $(i = 1, \ldots, n)$ identifiés avec les éléments de la base $\boldsymbol{p}$. Il ressort de la formule (5) qu'ils satisfont à la condition

$$(p^i, e_j) = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

On en déduit sans difficulté que pour $n = 3$ la base biorthogonale définie au § 3 du ch. I coïncide avec la base duale définie à la p. 224.

§ 2. Formes quadratiques

1. Formes bilinéaires. DÉFINITION. On appelle *fonction bilinéaire* ou *forme bilinéaire* sur un espace vectoriel $\mathscr{L}_n$ la fonction $\mathbf{b}$ de deux vecteurs de $\mathscr{L}_n$ vérifiant (pour tous vecteurs x, y et z et tout nombre α) les égalités

$$\left.\begin{aligned} \mathbf{b}(x+y,z) &= \mathbf{b}(x,z)+\mathbf{b}(y,z), \quad \mathbf{b}(\alpha x,y) = \alpha\mathbf{b}(x,y), \\ \mathbf{b}(x,y+z) &= \mathbf{b}(x,y)+\mathbf{b}(x,z), \quad \mathbf{b}(x,\alpha y) = \alpha\mathbf{b}(x,y). \end{aligned}\right\} \tag{1}$$

Choisissons dans l'espace $\mathscr{L}_n$ une base $\| e_1, \ldots, e_n \|$. Si $x = \sum \xi^i e_i$ et $y = \sum \eta^j e_j$, la valeur de la forme bilinéaire $\mathbf{b}$ sur les vecteurs x et y peut être calculée de la façon suivante :

$$\mathbf{b}(x,y) = \mathbf{b}\left(\sum_{i=1}^{n} \xi^i e_i, \sum_{j=1}^{n} \eta^j e_j\right) = \sum_{i,j} \xi^i \eta^j \mathbf{b}(e_i, e_j),$$

ou en définitive :

$$\mathbf{b}(x,y) = \sum_{i,j} \beta_{ij} \xi^i \eta^j. \tag{2}$$

Les n^2 nombres β_{ij}, valeurs de la forme bilinéaire sur tous les couples de vecteurs de base, s'appellent *coefficients de la forme bilinéaire* dans la base $\| e_1, \ldots, e_n \|$. On les écrit habituellement sous forme de matrice carrée d'ordre n :

$$B = \begin{Vmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ \beta_{n1} & \beta_{n2} & \ldots & \beta_{nn} \end{Vmatrix}.$$

Cette matrice est appelée *matrice de la forme bilinéaire* par rapport à la base donnée. On vérifie sans difficulté par multiplication des matrices qu'en représentation matricielle l'égalité (2) s'écrit sous la forme

$$\mathbf{b}(x,y) = {}^t\boldsymbol{\xi} B \boldsymbol{\eta}. \tag{3}$$

Avec le changement de base, la matrice associée à la forme bilinéaire varie évidemment. Cherchons la loi de cette variation. Supposons que les vecteurs $e_1', \ldots, e_n'$ de la nouvelle base s'expriment en fonction des vecteurs $e_1, \ldots, e_n$ de l'ancienne base par les égalités $e_i' = \sum \sigma_i^k e_k$, où σ_i^k désignent les éléments de la matrice de passage S. Les coefficients de la forme bili-

néaire $\mathbf{b}$ dans la base e' vérifient pour tous $i, j = 1, \ldots, n$

$$\mathbf{b}(e_i', e_j') = \mathbf{b}\left(\sum_{k=1}^{n} \sigma_i^k e_k, \sum_{l=1}^{n} \sigma_j^l e_l\right) = \sum_{k,l} \sigma_i^k \sigma_j^l \mathbf{b}(e_k, e_l),$$

ou

$$\beta_{ij}' = \sum_{k,l} \sigma_i^k \sigma_j^l \beta_{kl}. \tag{4}$$

Il est aisé de vérifier que l'égalité (4) est équivalente à l'égalité matricielle

$$B' = {}^tSBS, \tag{5}$$

où B' est la matrice de la forme bilinéaire par rapport à la base e'.

La forme bilinéaire $\mathbf{b}$ est dite *symétrique* si pour tous x et y on a l'égalité $\mathbf{b}(x, y) = \mathbf{b}(y, x)$.

Si la forme bilinéaire est symétrique, on a $\mathbf{b}(e_i, e_j) = \mathbf{b}(e_j, e_i)$ pour tous i et j et, partant, la matrice associée à la forme bilinéaire est symétrique. Inversement, supposons que la matrice de la forme bilinéaire est symétrique, c'est-à-dire que $B = {}^tB$. Alors, vu que la matrice de type (1, 1) ne varie pas par transposition, on a

$$\mathbf{b}(x, y) = {}^t({}^t\xi B \eta) = {}^t\eta\, {}^tB \xi = {}^t\eta B \xi = \mathbf{b}(y, x),$$

et, par suite, la forme bilinéaire $\mathbf{b}$ est symétrique. On a ainsi démontré la proposition suivante :

PROPOSITION 1. *Une forme bilinéaire est symétrique si et seulement si sa matrice est symétrique (quelle que soit la base).*

2. Autre aspect de la forme bilinéaire. Considérons une forme bilinéaire $\mathbf{b}(x, y)$ définie sur l'espace vectoriel $\mathcal{L}_n$ et fixons un vecteur quelconque y de $\mathcal{L}_n$. $\mathbf{b}(x, y)$ est alors une fonction linéaire de x. Pour le souligner, on peut désigner $\mathbf{b}(x, y)$ par $\mathbf{b}_y(x)$. Ainsi, à chaque vecteur y de $\mathcal{L}_n$ est associée une fonction linéaire sur $\mathcal{L}_n$, autrement dit on a une application $\mathbf{B} : \mathcal{L}_n \to \mathcal{L}_n^*$. On voit aisément que l'application $\mathbf{B}$ est linéaire. En effet, considérons l'image $\mathbf{B}(x + y)$ de la somme de deux vecteurs x et y. C'est une fonction linéaire $\mathbf{b}_{x+y}$ qui fait correspondre à tout vecteur z le nombre $\mathbf{b}(z, x + y)$. La fonction $\mathbf{b}$ étant linéaire par rapport au deuxième argument, on a $\mathbf{b}(z, x + y) = \mathbf{b}(z, x) + \mathbf{b}(z, y)$. D'où $\mathbf{b}_{x+y} = \mathbf{b}_x + \mathbf{b}_y$. Donc, $\mathbf{B}(x + y) = \mathbf{B}(x) + \mathbf{B}(y)$. De façon analogue on démontre que $\mathbf{B}(\alpha x) = \alpha \mathbf{B}(x)$.

Inversement, soit donnée une application linéaire $\mathbf{B} : \mathcal{L}_n \to \mathcal{L}_n^*$ qui fait correspondre à tout vecteur y de $\mathcal{L}_n$ une fonction linéaire $\mathbf{b}_y$ de $\mathcal{L}_n^*$. Donc, à tout couple de vecteurs x et y de $\mathcal{L}_n$ on peut faire correspondre un nombre $\mathbf{b}_y(x)$ qu'on est en droit de noter $\mathbf{b}(x, y)$. La fonction $\mathbf{b}(x, y)$ est linéaire par rapport au premier argument, vu que $\mathbf{b}_y$ est une fonction linéaire; elle est aussi linéaire par rapport au deuxième argument, vu que $\mathbf{B}$ est une application linéaire.

Ainsi donc, la forme bilinéaire sur $\mathcal{L}_n$ peut être définie comme une application linéaire de $\mathcal{L}_n$ dans $\mathcal{L}_n^*$.

Soit e une base dans $\mathcal{L}_n$ et p la base duale dans $\mathcal{L}^*$. Cherchons la matrice de l'application linéaire $\mathbf{B}$ correspondant à la forme bilinéaire $\mathbf{b}$. L'image $\mathbf{B}(e_i)$ du vecteur de base e_i est la fonction linéaire $\mathbf{b}_{e_i}(x)$. Ses coefficients sont $\mathbf{b}(e_1, e_i), \ldots, \mathbf{b}(e_n, e_i)$. Les mêmes nombres for-

ment la i-ième colonne de la matrice cherchée, vu que ce sont les coordonnées de $\mathbf{B}(e_i)$ par rapport à la base p. On constate que la matrice de l'application $\mathbf{B}$ par rapport aux bases e et p coïncide avec la matrice B de la forme bilinéaire $\mathbf{b}$ par rapport à la base e.

Il découle de ce qui vient d'être dit que la loi de variation de la matrice associée à la forme bilinéaire (5) est un simple corollaire de la formule (6) du § 3, ch. VI (transformation de la matrice d'une application linéaire) et de l'égalité (8) du § 1.

Le théorème 2 du § 3, ch. VI, ne se rapporte pas aux formes bilinéaires car les bases e et p doivent être biorthogonales et ne peuvent donc être choisies de façon arbitraire.

Si la forme bilinéaire n'est pas symétrique, en fixant le premier argument au lieu du second on obtient une autre application ${}^t\mathbf{B} : \mathscr{L} \to \mathscr{L}^*$. Sa matrice par rapport aux bases biorthogonales e et p est égale à la matrice transposée ${}^tB$ de la forme bilinéaire. Nous proposons au lecteur de le vérifier à titre d'exercice.

3. Formes quadratiques. On passe maintenant à l'étude d'une classe importante de fonctions définies sur un espace vectoriel, qui sont étroitement liées aux formes bilinéaires.

DÉFINITION. On appelle *forme quadratique* la fonction $\mathbf{k}$ sur l'espace vectoriel $\mathscr{L}_n$ dont la valeur sur tout vecteur x est définie par l'égalité $\mathbf{k}(x) = \mathbf{b}(x, x)$, où $\mathbf{b}$ est une forme bilinéaire symétrique sur $\mathscr{L}_n$.

La forme quadratique $\mathbf{k}$ définit de façon univoque la forme bilinéaire symétrique correspondante $\mathbf{b}$. En effet, soient x et y des vecteurs quelconques. Considérons la valeur de la forme quadratique sur le vecteur $x + y$

$$\mathbf{k}(x + y) = \mathbf{b}(x + y, x + y) = \mathbf{b}(x, x) + \mathbf{b}(x, y) + \mathbf{b}(y, x) + \mathbf{b}(y, y).$$

La forme bilinéaire étant symétrique, on obtient

$$\mathbf{b}(x, y) = \frac{1}{2}[\mathbf{k}(x + y) - \mathbf{k}(x) - \mathbf{k}(y)],$$

et, par suite, la valeur de $\mathbf{b}$ sur tout couple de vecteurs s'exprime en fonction des valeurs de $\mathbf{k}$.

On appelle *matrice de la forme quadratique* $\mathbf{k}$ la matrice de la forme bilinéaire symétrique correspondante $\mathbf{b}$.

Selon (2), la valeur $\mathbf{k}(x)$ de la forme quadratique $\mathbf{k}$ s'écrit au moyen des coordonnées du vecteur x par rapport à une base de la façon suivante :

$$\mathbf{k}(x) = \sum_{i,j} \beta_{ij}\xi^i\xi^j \tag{6}$$

ou sous forme matricielle

$$\mathbf{k}(x) = {}^t\boldsymbol{\xi}B\boldsymbol{\xi}. \tag{7}$$

Le second membre de (6) est un polynôme homogène du second degré en $\xi^1, \ldots, \xi^n$. Son expression contient des termes semblables. A savoir, pour $i \neq j$, $\beta_{ij}\xi^i\xi^j$ et $\beta_{ji}\xi^j\xi^i$ coïncident. Il en ressort qu'après la réduction des termes semblables, (6) prend la forme

$$\mathbf{k}(x) = \beta_{11}(\xi^1)^2 + 2\beta_{12}\xi^1\xi^2 + \beta_{22}(\xi^2)^2 + 2\beta_{13}\xi^1\xi^3 + \ldots \tag{8}$$

THÉORÈME 1. *Pour toute forme quadratique* $\mathbf{k}$ *il existe une base dans laquelle*

$$\mathbf{k}(x) = \sum_{i=1}^{n} \varepsilon_i(\xi^i)^2, \tag{9}$$

c'est-à-dire que la matrice associée à la forme quadratique est diagonale.

La relation (9) est appelée *décomposition en carrés* de la forme quadratique.

DÉMONSTRATION. Considérons une forme quadratique $\mathbf{k}$ et notons B sa matrice par rapport à une base de départ. On appliquera à la matrice B une suite de transformations élémentaires qu'on partagera, pour commodité de description, en étapes. A la première étape deux cas peuvent se présenter :

1) Cas général : $\beta_{11} \neq 0$. Si cette condition est remplie, retranchons la première ligne multipliée par des facteurs convenables (β_{1i}/β_{11} pour la i-ième ligne) de toutes les lignes situées au-dessous d'elle. De façon analogue, retranchons la première colonne multipliée par les mêmes facteurs de toutes les colonnes situées à droite d'elle. Les facteurs sont choisis de manière que la matrice B se transforme en matrice B_1

$$\left\| \begin{matrix} \varepsilon_1 & 0 & \dots & 0 \\ 0 & & & \\ \dots & & C_1 & \\ 0 & & & \end{matrix} \right\|, \tag{10}$$

où C_1 est une matrice carrée symétrique d'ordre $n - 1$.

2) Cas particulier : $\beta_{11} = 0$. Deux possibilités apparaissent : a) $\beta_{1i} = 0$ pour tous les $i = 2, \dots, n$. La matrice est déjà de la forme (10). b) Il existe un i pour lequel $\beta_{1i} \neq 0$. On effectue alors une transformation supplémentaire : si $\beta_{ii} \neq 0$, on permute la i-ième ligne avec la première, ainsi que la i-ième colonne avec la première. On a alors $\beta'_{11} = \beta_{ii} \neq 0$. Mais si $\beta_{ii} = 0$ on ajoute la i-ième ligne à la première, et la i-ième colonne à la première. Dans la matrice transformée, $\beta'_{11} = 2\beta_{1i} \neq 0$, de sorte qu'après une transformation supplémentaire la matrice se réduit à la forme (10) comme dans le cas général.

Supposons qu'après k étapes on ait obtenu la matrice B_k de la forme

$$\left\| \begin{array}{ccc|c} \varepsilon_1 & & & \\ & \ddots & & O \\ & & \varepsilon_k & \\ \hline & O & & C_k \end{array} \right\|. \tag{11}$$

C_k est ici une matrice symétrique d'ordre $n - k$, et $\varepsilon_1, \dots, \varepsilon_k$ sont les éléments supérieurs gauches des matrices C_i obtenues aux étapes précédentes.

La $(k+1)$-ième étape consiste dans l'application à la matrice B_k des transformations élémentaires qui sont équivalentes aux transformations de la première étape appliquées à la matrice C_k, et qui n'affectent pas les k premières lignes et les k premières colonnes de la matrice B_k. On obtient en définitive la matrice B_{k+1} qui est de même type que la matrice B_k.

Après $n-1$ étapes la matrice C_{n-1} est d'ordre 1 et n'a plus besoin d'être transformée. En définitive, la matrice B se transforme en matrice diagonale

$$B' = \begin{Vmatrix} \varepsilon_1 & & \\ & \ddots & \\ & & \varepsilon_n \end{Vmatrix}.$$

Il va de soi que si la matrice de départ ou l'une des matrices C_k est nulle, les transformations suivantes deviennent inutiles, car la matrice est déjà diagonale. Cette situation est équivalente à la réalisation du cas particulier à toutes les étapes ultérieures.

Il est important de signaler qu'après chaque transformation élémentaire des lignes on a effectué la même transformation des colonnes. Si la transformation élémentaire des colonnes de la matrice B est équivalente à la multiplication à droite de la matrice B par la matrice S_α (comp. point 4, § 6, ch. V), la même transformation des lignes est équivalente à la multiplication à gauche de la matrice B par la matrice ${}^tS_\alpha$.

A la suite de toutes ces transformations élémentaires on obtient la matrice $B' = {}^tSBS$, où $S = S_1 \ldots S_N$ est le produit de toutes les matrices correspondant aux transformations des colonnes.

On a ainsi démontré que la matrice B' est la matrice de la forme quadratique $\mathbf{k}$ par rapport à la base e' qui est liée à la base de départ e par la matrice de passage S. Le théorème est démontré.

Pour décomposer la forme quadratique en carrés on peut se servir de la *méthode de Lagrange*. Montrons-le sur un exemple. Soit une forme quadratique de type (8)

$$\mathbf{k}(x) = 2(\xi^1)^2 + 4\xi^1\xi^2 + 3(\xi^2)^2 + 4\xi^2\xi^3 + 5(\xi^3)^2.$$

Le coefficient de $(\xi^1)^2$ étant différent de zéro, groupons tous les termes contenant ξ^1 :

$$2[(\xi^1)^2 + 2\xi^1\xi^2] + 3(\xi^2)^2 + 4\xi^2\xi^3 + 5(\xi^3)^2.$$

Complétons maintenant l'expression entre crochets pour avoir le carré de la somme. A cet effet, ajoutons et retranchons $2(\xi^2)^2$:

$$2[(\xi^1)^2 + 2\xi^1\xi^2 + (\xi^2)^2] - 2(\xi^2)^2 + 3(\xi^2)^2 + 4\xi^2\xi^3 + 5(\xi^3)^2.$$

Maintenant on peut écrire la forme quadratique de la façon suivante :

$$\mathbf{k}(x) = 2(\xi^1 + \xi^2)^2 + \mathbf{k}'(x),$$

où $\mathbf{k}'$ est une forme quadratique dont la valeur ne dépend que de ξ^2 et ξ^3 :

$$\mathbf{k}'(x) = (\xi^2)^2 + 4\xi^2\xi^3 + 5(\xi^3)^2.$$

On peut y appliquer le même procédé :

$$\mathbf{k}'(x) = (\xi^2 + 2\xi^3)^2 + (\xi^3)^2.$$

Maintenant $\mathbf{k}(x)$ prend la forme de

$$2(\tilde{\xi}^1)^2 + (\tilde{\xi}^2)^2 + (\tilde{\xi}^3)^2,$$

où

$$\tilde{\xi}^1 = \xi^1 + \xi^2, \quad \tilde{\xi}^2 = \xi^2 + 2\xi^3, \quad \tilde{\xi}^3 = \xi^3.$$

Les dernières formules définissent la transformation des coordonnées d'un vecteur lorsqu'on passe à la base dans laquelle la forme quadratique se décompose en carrés.

Dans la démonstration du théorème 1, on a proposé une suite déterminée de transformations élémentaires pour réduire la forme quadratique à la somme des carrés. La méthode de Lagrange ne diffère du procédé proposé que par la forme d'écriture. Il est utile de savoir que pour réduire une matrice à la forme diagonale, on peut utiliser toute autre suite de transformations élémentaires, mais à une seule condition importante : après chaque transformation élémentaire des lignes il faut procéder à la même transformation élémentaire des colonnes.

La décomposition en carrés de la forme quadratique dans un espace réel sera appelée *forme canonique* si les nombres ε_k ne peuvent prendre que les valeurs 1, -1 et 0. Dans un espace complexe, la décomposition en carrés de la forme quadratique est canonique si ε_k ne sont égaux qu'à 1 ou 0.

THÉORÈME 2. *Pour toute forme quadratique il existe une base par rapport à laquelle elle est de la forme canonique.*

Pour le démontrer, décomposons d'abord la forme quadratique en carrés, après quoi faisons la transformation suivante. Si l'un quelconque des éléments diagonaux ε_k est différent de zéro, multiplions le k-ième vecteur de base par $(\varepsilon_k)^{-1/2}$ ou par $|\varepsilon_k|^{-1/2}$ selon que l'espace est complexe ou réel. On voit facilement que cela correspond à la multiplication de la k-ième ligne et de la k-ième colonne de la matrice de la forme quadratique par le même facteur. En le faisant pour tous les k tels que $\varepsilon_k \neq 0$, on réduit la forme quadratique à sa forme canonique.

4. Rang et indice de la forme quadratique. Il existe plusieurs bases dans lesquelles la forme quadratique donnée prend une forme canonique. Pour chacune de ces bases, les coefficients ε_k auraient dû en général être différents. Or il s'avère justement qu'ils sont les mêmes (à l'ordre près), quel que soit le procédé appliqué pour réduire la forme quadratique à sa forme canonique.

Commençons par la proposition auxiliaire importante suivante.

PROPOSITION 2. *Si* det $A \neq 0$ *et les produits* AB *et* CA *sont définis, on a* Rg AB = Rg B *et* Rg CA = Rg C.

En effet, selon la proposition 5 du § 6, ch. V, on a simultanément Rg B = Rg $A^{-1}(AB) \leqslant$ Rg AB et Rg $AB \leqslant$ Rg B, d'où Rg B = Rg AB. La proposition se démontre de façon analogue si le facteur A est à droite.

THÉORÈME 3. *Le rang de la matrice associée à la forme quadratique est indépendant de la base.*

En effet, selon la formule (5), les matrices K et K' de la forme quadratique **k** rapportée aux bases $\boldsymbol{e}$ et $\boldsymbol{e}'$ sont liées par l'égalité $K' = {}^tSKS$, où det $S \neq 0$. D'où en vertu de la proposition 2, Rg K' = Rg KS = Rg K. Si la forme quadratique est de la forme canonique, le rang de sa matrice est égal au nombre des coefficients ε_i différents de zéro. Or, ce nombre ne dépend pas de la base.

DÉFINITION. On appelle *rang* de la forme quadratique **k** le nombre des coefficients ε_i non nuls de sa forme canonique.

On a vu que le rang de **k** est égal au rang de sa matrice dans une base quelconque.

Signalons que la représentation « géométrique » du rang de la forme quadratique peut être obtenue en interprétant la forme bilinéaire symétrique correspondante comme une application $\mathscr{L}_n \to \mathscr{L}_n^*$.

Dans un espace complexe, toutes les formes quadratiques d'un même rang r se réduisent (chacune dans sa base) à la même forme canonique $(\xi^1)^2 + \ldots + (\xi^r)^2$.

Considérons maintenant la forme quadratique **k** sur l'espace réel $\mathscr{L}_n$.

DÉFINITION. On dit que la forme quadratique **k** est *définie positive sur le sous-espace* $\mathscr{L}'$ de l'espace $\mathscr{L}_n$ si $\mathbf{k}(x) > 0$ pour tout vecteur x non nul de $\mathscr{L}'$. D'une façon analogue, **k** est *définie négative sur* $\mathscr{L}'$ si $\mathbf{k}(x) < 0$ pour tout x non nul de $\mathscr{L}'$.

Si $\mathbf{k}(x) > 0$ ou $\mathbf{k}(x) < 0$ pour tout $x \neq 0$ de $\mathscr{L}_n$, on dit que **k** est respectivement une forme quadratique *définie positive* ou une forme quadratique *définie négative*.

Les formes quadratiques qui pour tout x vérifient les inégalités $\mathbf{k}(x) \geqslant 0$ ou $\mathbf{k}(x) \leqslant 0$ sont dites respectivement *semi-définies positives* ou *négatives*.

Il est commode de poser que sur le sous-espace nul, toute forme quadratique est à la fois définie positive et négative. En vertu de cette convention il existe toujours un sous-espace (au moins nul) sur lequel la forme quadratique est définie négative, et l'on est en mesure de choisir entre ces sous-espaces celui qui est de dimension maximale.

DÉFINITION. Soit $\mathscr{L}^{(-)}$ un sous-espace de dimension maximale parmi tous les sous-espaces sur lesquels la forme quadratique $\mathbf{k}$ est définie négative. La dimension de $\mathscr{L}^{(-)}$ est appelée *indice négatif* de la forme quadratique $\mathbf{k}$.

THÉORÈME 4 (loi d'inertie des formes quadratiques). *Etant donné la forme quadratique* $\mathbf{k}$, *le nombre des coefficients* ε_i *négatifs (positifs) de la forme canonique ne dépend pas du choix de la base dans laquelle elle est réduite à la forme canonique.*

On démontrera une affirmation équivalente : si dans une base quelconque la forme quadratique $\mathbf{k}$ est réduite à la forme canonique, le nombre des coefficients négatifs coïncide avec l'indice négatif de $\mathbf{k}$. Etant donné que le nombre total des coefficients positifs et négatifs est égal au rang, il ressort de la proposition précédente que le nombre des coefficients positifs est aussi indépendant de la base.

En effet, supposons que dans la base $\|e_1, \ldots, e_n\|$ la forme quadratique d'indice négatif s est de la forme canonique

$$-(\xi^1)^2 - \ldots - (\xi^j)^2 + (\xi^{j+1})^2 + \ldots + (\xi^r)^2.$$

Désignons par $\mathscr{L}_1$ l'enveloppe linéaire des vecteurs $e_1, \ldots, e_j$ et par $\mathscr{L}_2$ celle des vecteurs $e_{j+1}, \ldots, e_n$. Pour un vecteur x de $\mathscr{L}_1$ on a $\xi^{j+1} = \ldots = \xi^n = 0$, de sorte que $\mathbf{k}(x) = -(\xi^1)^2 - \ldots - (\xi^j)^2 < 0$ si $x \neq 0$. Donc, la forme $\mathbf{k}$ est définie négative sur le sous-espace $\mathscr{L}_1$ de dimension j, et $s \geqslant j$.

Admettons maintenant que $s > j$ et qu'il existe un sous-espace $\mathscr{L}^{(-)}$ de dimension s sur lequel $\mathbf{k}$ est définie négative. Les composantes de tout vecteur x de $\mathscr{L}_2$ vérifient les égalités $\xi^1 = \ldots = \xi^j = 0$ et, par suite $\mathbf{k}(x) \geqslant 0$ pour tout x de $\mathscr{L}_2$. La dimension de $\mathscr{L}_2$ est $n - j$, et la somme des dimensions de $\mathscr{L}_2$ et de $\mathscr{L}^{(-)}$ est strictement supérieure à n. Selon le théorème 1 du § 2, ch. VI, $\mathscr{L}_2$ et $\mathscr{L}^{(-)}$ ont une intersection non nulle. Pour un vecteur z non nul de $\mathscr{L}_2 \cap \mathscr{L}^{(-)}$ on aurait dû avoir simultanément $\mathbf{k}(z) < 0$ et $\mathbf{k}(z) \geqslant 0$. Il découle de la contradiction obtenue que $s = j$. Le théorème est démontré.

Les formes quadratiques définies positives sont de rang n et d'indice négatif 0 et se réduisent à la forme canonique

$$(\xi^1)^2 + \ldots + (\xi^n)^2. \tag{12}$$

Les formes quadratiques définies négatives sont de rang n et d'indice négatif n. Elles se réduisent à la forme

$$-(\xi^1)^2 - \ldots - (\xi^n)^2.$$

Les formes semi-définies positives et négatives de rang r se réduisent respectivement aux formes canoniques suivantes :

$$(\xi^1)^2 + \ldots + (\xi^r)^2, \quad -(\xi^1)^2 - \ldots - (\xi^r)^2.$$

On a vu qu'à toute forme quadratique sur un espace réel on peut associer deux nombres : son rang et son indice négatif. Ces deux nombres la caractérisent au sens que toutes les formes quadratiques ayant même rang et même indice négatif se réduisent (chacune dans sa base) à la même forme canonique. Au lieu du rang et de l'indice négatif, la forme quadratique peut être caractérisée par tout autre couple de grandeurs permettant de trouver ces nombres. Par exemple, au lieu du rang on se donne souvent l'indice positif qui est égal au nombre des coefficients positifs dans la forme canonique. Avec l'indice négatif, nombre des coefficients négatifs, il caractérise la forme quadratique. Dans le § 2 du ch. IX on caractérisera la forme quadratique par le rang et par la différence des indices positif et négatif, appelée *signature* de la forme quadratique.

Il est utile de savoir déterminer si la forme quadratique donnée est définie positive, et cela sans la rendre canonique. On peut le faire en se servant du théorème suivant appelé *loi de Sylvester*.

THÉORÈME 5. *Pour qu'une forme quadratique soit définie positive, il faut et il suffit que les mineurs de sa matrice vérifient les inégalités*

$$\begin{Vmatrix} \beta_{11} & \dots & \beta_{1k} \\ \dots & \dots & \dots \\ \beta_{k1} & \dots & \beta_{kk} \end{Vmatrix} > 0 \tag{13}$$

pour tous $k = 1, \dots, n$.

Les mineurs de la forme (13) sont appelés *mineurs principaux* de la matrice.

Pour le démontrer, reprenons les transformations de la matrice associée à la forme quadratique, qui ont été utilisées dans la démonstration du théorème 1.

1° La condition est *nécessaire*. Si la forme quadratique $\mathbf{k}$ est définie positive, les éléments diagonaux de sa matrice par rapport à toute base satisfont à la condition

$$\beta_{ii} = \mathbf{k}(e_i) > 0,$$

et, par suite, en réduisant la matrice à la forme diagonale on ne se heurte pas au cas particulier. Dans le cas général, à toute ligne on ne peut ajouter que la ligne située au-dessus d'elle et à toute colonne, que la colonne située à gauche d'elle. Les mineurs principaux de la matrice ne varient pas par ces transformations. Or les mineurs principaux de la matrice diagonale associée à la forme quadratique définie positive sont positifs. Par suite, les mineurs principaux de la matrice initiale sont aussi positifs.

2° La condition est *suffisante*. Supposons que tous les mineurs principaux de la matrice B sont positifs. En particulier, $M_1 = \beta_{11} > 0$, et les transformations de la première étape conduisent la matrice à la forme (10)

avec $\varepsilon_1 > 0$. Admettons qu'après k étapes on obtient la matrice B_k avec $\varepsilon_1, \ldots, \varepsilon_k$ strictement positifs, sans qu'apparaisse le cas particulier. Alors pour l'élément supérieur gauche de la matrice C_k on a $\varepsilon_{k+1} = M_{k+1}/M_k$ car les mineurs principaux n'ont pas varié. Il s'ensuit que $\varepsilon_{k+1} > 0$. A l'étape suivante des transformations on a le cas général, et la matrice obtenue B_{k+1} contient les éléments $\varepsilon_1, \ldots, \varepsilon_{k+1}$ strictement positifs. En procédant à ce raisonnement pour tous les k de 2 à n, on finit par démontrer le théorème.

§ 3. Formes quadratiques et produit scalaire

Si dans un espace vectoriel réel est défini le produit scalaire (autrement dit l'espace est euclidien), on peut à chaque forme bilinéaire, indépendamment du choix de la base, faire correspondre une transformation linéaire.

DÉFINITION. La transformation linéaire $\mathbf{A}$ de l'espace euclidien $\mathcal{E}_n$ est dite *associée* à la forme bilinéaire $\mathbf{b}$ si pour tous vecteurs x et y de $\mathcal{E}_n$ est vérifiée l'égalité

$$\mathbf{b}(x, y) = (x, \mathbf{A}(y)). \qquad (1)$$

PROPOSITION 1. *Pour toute forme bilinéaire il existe une transformation associée et une seule.*

Pour le démontrer, admettons d'abord que la transformation associée existe pour $\mathbf{b}$. Notons B et A les matrices de $\mathbf{b}$ et de $\mathbf{A}$ par rapport à une base e et écrivons l'égalité (1) sous forme matricielle :

$${}^t\boldsymbol{\xi} B \boldsymbol{\eta} = {}^t\boldsymbol{\xi} \Gamma A \boldsymbol{\eta},$$

où Γ est la matrice de Gram de la base e. On peut donner à la dernière égalité la forme ${}^t\boldsymbol{\xi}(B - \Gamma A)\boldsymbol{\eta} = 0$.

Supposons que les vecteurs arbitraires x et y sont égaux aux vecteurs de base e_i et e_j. Alors leurs colonnes de coordonnées $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ sont respectivement égales à la i-ième et à la j-ième colonne de la matrice unité. Il s'ensuit que l'élément de la i-ième ligne et de la j-ième colonne de la matrice $B - \Gamma A$ est nul (comp. exemple 4, p. 160). Etant donné que i et j sont arbitraires, on voit que $B - \Gamma A = O$. On peut donc exprimer la matrice A au moyen de la matrice B :

$$A = \Gamma^{-1} B. \qquad (2)$$

Cela signifie que la forme bilinéaire ne peut avoir plus d'une transformation associée : si elle existe, sa matrice est $\Gamma^{-1}B$.

On démontre maintenant sans difficulté l'existence de la transformation associée. Il suffit pour cela de vérifier que la transformation de matrice (2) est la transformation associée. En effet, il ressort de (2) que

$B = \Gamma A$ et, par suite, pour toutes matrices-colonnes $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ on a ${}^t\boldsymbol{\xi}B\boldsymbol{\eta} = {}^t\boldsymbol{\xi}\Gamma A\boldsymbol{\eta}$. L'égalité (1) est donc vérifiée par tous vecteurs x et y de $\mathscr{E}_n$. La proposition est démontrée.

Signalons que dans le cas d'une base orthonormée, la relation entre les matrices de la forme bilinéaire et de la transformation associée est particulièrement simple:

$$B = A.$$

Ce résultat et la proposition 2 du § 2, ch. VI, entraînent la

PROPOSITION 2. *La transformation associée à une forme bilinéaire est symétrique si et seulement si cette forme est symétrique.*

Si la transformation linéaire est associée à une forme bilinéaire symétrique **b**, on dit qu'elle est aussi associée à la forme quadratique définie par **b**. Ainsi, à chaque forme quadratique est associée une transformation symétrique.

Le lien étroit entre les formes quadratiques et les transformations symétriques permet de démontrer le théorème important suivant.

THÉORÈME 1. *Dans un espace euclidien, il existe pour chaque forme quadratique une base orthonormée par rapport à laquelle elle se décompose en carrés.*

Le théorème est presque évident : la base dont on affirme l'existence est une base orthonormée constituée des vecteurs propres de la transformation linéaire associée à la forme quadratique. Dans cette base on a $B = A$ et A est une matrice diagonale.

THÉORÈME 2. *L'espace euclidien peut être défini comme un espace vectoriel réel muni d'une forme quadratique définie positive.*

En effet, faisons correspondre à chaque vecteur de l'espace euclidien $\mathscr{E}_n$ son carré scalaire. On obtient une fonction définie par $\mathbf{k}(x) = (x, x)$. Si Γ est la matrice de Gram d'une base de $\mathscr{E}_n$ et $\boldsymbol{\xi}$ la colonne de coordonnées d'un vecteur x dans cette base, la fonction **k** se définit par la formule $\mathbf{k}(x) = {}^t\boldsymbol{\xi}\Gamma\boldsymbol{\xi}$. C'est donc une forme quadratique. En vertu de l'axiome 4° de l'espace euclidien, $\mathbf{k}(x) > 0$ pour $x \neq 0$, de sorte que la forme **k** est définie positive.

Inversement, soit dans l'espace vectoriel réel $\mathscr{L}_n$ une forme quadratique définie positive **k**. Elle permet de définir une forme bilinéaire symétrique **b** et une seule. Introduisons dans $\mathscr{L}_n$ le produit scalaire en posant $(x, y) = \mathbf{b}(x, y)$. Les axiomes 2° et 3° sont vérifiés car **b** est une forme bilinéaire. L'axiome 1° se vérifie vu que **b** est symétrique. La forme **k** étant définie positive, l'axiome 4° est vérifié. Le théorème est démontré.

Notons que (x, y) s'exprime dans la base $\boldsymbol{e}$ par l'égalité $(x, y) = {}^t\boldsymbol{\xi}K\boldsymbol{\eta}$,

où ${}^t\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ sont les colonnes de coordonnées des vecteurs x et y, et K la matrice de la forme quadratique. K est donc la matrice de Gram de la base e. Il ressort de la démonstration du théorème 2 que la forme quadratique doit être définie positive pour que soit vérifié l'axiome 4°.

On étudie aussi des espaces dans lesquels le produit scalaire est défini à l'aide d'une forme quadratique quelconque. Dans ces espaces, il existe des vecteurs dont le carré scalaire est négatif, ce qui rend la géométrie profondément différente de l'euclidienne. Si le produit scalaire est défini par une forme quadratique de rang égal à la dimension de l'espace, on dit que c'est un espace *pseudo-euclidien*. Parmi les espaces pseudo-euclidiens, un rôle important en physique mathématique est joué par l'espace à quatre dimensions, dans lequel est définie la forme quadratique $-(\xi^1)^2 - (\xi^2)^2 - (\xi^3)^2 + (\xi^4)^2$. C'est l'espace dit *de Minkowski*.

Utilisons le théorème 2 pour démontrer le théorème suivant.

THÉORÈME 3. *Soient dans un espace vectoriel $\mathscr{L}_n$ deux formes quadratiques* $\mathbf{k}$ *et* $\mathbf{h}$, *dont* $\mathbf{h}$ *est définie positive. Il existe alors dans $\mathscr{L}_n$ une base par rapport à laquelle les deux formes se décomposent en carrés* (*la forme* $\mathbf{h}$ *étant d'ailleurs de la forme canonique*).

Pour le démontrer, introduisons dans $\mathscr{L}_n$ le produit scalaire à l'aide de la forme définie positive $\mathbf{h}$. Toute base dans laquelle $\mathbf{h}$ est de la forme (12) du § 2 est une base orthonormée par rapport à ce produit scalaire, vu que sa matrice de Gram est une matrice unité. Selon le théorème 1 il existe pour la forme $\mathbf{k}$ une base orthonormée dans laquelle elle se décompose en carrés. C'est justement la base dont on démontre l'existence.

REMARQUE. Si $\mathscr{L}_n$ est euclidien, le théorème 3 demeure évidemment vrai. On s'abstrait du produit scalaire défini auparavant et on introduit un nouveau produit scalaire par l'intermédiaire de la forme $\mathbf{h}$.

La base dans laquelle $\mathbf{k}$ et $\mathbf{h}$ se décomposent en carrés ne sera pas, dans le cas général, orthonormée par rapport à l'ancien produit scalaire.

En pratique, pour décomposer en carrés les deux formes quadratiques, on construit d'abord une base dans laquelle $\mathbf{h}$ se décompose en carrés et l'on recherche la matrice K' de la forme $\mathbf{k}$ dans cette base. On passe ainsi à une base orthonormée par rapport au produit scalaire auxiliaire introduit lors de la démonstration du théorème. La transformation linéaire dont la matrice par rapport à la base trouvée est aussi K', est la transformation associée à la forme $\mathbf{k}$. Il faut rechercher les vecteurs propres de cette transformation, les orthogonaliser et les normer en calculant le produit scalaire suivant la formule (11) du § 1, ch. VII. On obtient ainsi la base des vecteurs propres, orthonormés par rapport au produit scalaire auxiliaire. C'est justement la base recherchée. La matrice de la forme $\mathbf{h}$ y est une matrice unité et la matrice K'' de la forme $\mathbf{k}$ est diagonale, ses éléments diagonaux étant égaux aux racines du polynôme caractéristique de la matrice K'.

La décomposition en carrés des deux formes quadratiques peut aussi être réalisée d'une autre façon. Soient K et H les matrices des formes quadratiques dans la base initiale $\boldsymbol{e}$. Comme il a été signalé dans la démonstration du théorème 2, H est la matrice de Gram de la base $\boldsymbol{e}$ pour le produit scalaire auxiliaire. La matrice $A = H^{-1}K$ est la matrice de la transformation linéaire associée à la forme $\mathbf{k}$. L'équation caractéristique de la transformation est de la forme $\det (H^{-1}K - \lambda E) = 0$. Vu que $H^{-1}K - \lambda E = H^{-1}(K - \lambda H)$ et $\det H^{-1} \neq 0$, l'équation caractéristique possède les mêmes racines que l'équation

$$\det (K - \lambda H) = 0. \tag{3}$$

Pour chacune de ces racines, le système d'équations linéaires $(H^{-1}K - \lambda E)\boldsymbol{\xi} = \boldsymbol{o}$ servant à la recherche des vecteurs propres est équivalent au système

$$(K - \lambda H)\boldsymbol{\xi} = \boldsymbol{o}.$$

Pour chaque λ, le système fondamental des solutions d'un tel système doit être orthogonalisé et normé à l'aide du produit scalaire défini par la formule (13) du § 1, ch. VII, avec matrice de Gram H. On construit ainsi la base $\tilde{\boldsymbol{e}}$. Vu qu'elle est orthonormée par rapport au produit scalaire auxiliaire, la matrice de $\mathbf{h}$ dans la base $\tilde{\boldsymbol{e}}$ est une matrice unité. Cette base étant composée des vecteurs propres de la transformation associée à la forme $\mathbf{k}$, la matrice de $\mathbf{k}$ dans $\tilde{\boldsymbol{e}}$ est une matrice diagonale avec racines de l'équation (3) sur la diagonale.

§ 4. Formes hermitiennes

Les formes quadratiques définies sur les espaces unitaires ne vérifient pas les théorèmes du § 3. Ces théorèmes sont pourtant valables pour d'autres fonctions appelées *formes hermitiennes*. Limitons-nous aux énoncés de ces théorèmes car les démonstrations ne diffèrent presque pas de celles des théorèmes correspondants du § 3.

La fonction $\mathbf{b}$ de deux vecteurs, définie sur un espace vectoriel complexe $\mathscr{L}_n$ est appelée *forme bilinéaire hermitienne* ou *forme sesquilinéaire* si pour tous vecteurs x, y et z et tout nombre complexe α sont vérifiées les égalités

$$\mathbf{b}(x + y, z) = \mathbf{b}(x, y) + \mathbf{b}(y, z), \quad \mathbf{b}(\alpha x, y) = \alpha\mathbf{b}(x, y),$$
$$\mathbf{b}(x, y + z) = \mathbf{b}(x, y) + \mathbf{b}(x, z), \quad \mathbf{b}(x, \alpha y) = \bar{\alpha}\mathbf{b}(x, y).$$

Cette définition diffère de celle de la forme bilinéaire par le fait que si l'on multiplie le second argument par le nombre α, on doit multiplier la valeur de la forme bilinéaire hermitienne par le conjugué complexe $\bar{\alpha}$.

Si l'on a choisi une base dans $\mathcal{L}_n$, la valeur de **b** sur le couple de vecteurs (x, y) peut être écrite sous la forme

$$\mathbf{b}(x, y) = \sum_{i,j} \beta_{ij} \xi^i \bar{\eta}^j, \quad \text{ou} \quad \mathbf{b}(x, y) = {}^t\boldsymbol{\xi} B \bar{\boldsymbol{\eta}},$$

où $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ sont les colonnes de coordonnées des vecteurs x et y. La matrice $B = \|\beta_{ij}\|$ est appelée *matrice de la forme bilinéaire hermitienne*. Lorsqu'on change de base, elle se transforme suivant la loi ${}^tSB\bar{S}$.

La forme bilinéaire hermitienne est dite *symétrique* si $\mathbf{b}(x, y) = \overline{\mathbf{b}(y, x)}$. Cette condition est équivalente au fait que sa matrice est hermitienne.

La fonction **k** sur $\mathcal{L}_n$ s'appelle *forme quadratique hermitienne* ou tout simplement *forme hermitienne* si $\mathbf{k}(x) = \mathbf{b}(x, x)$ pour une forme bilinéaire hermitienne symétrique **b**. La forme **b** se définit univoquement par **k**.

Considérons une forme hermitienne **k** sur l'espace unitaire $\mathcal{U}_n$. La transformation linéaire **A** dans $\mathcal{U}_n$ est dite *associée* à la forme **k** si elle est auto-adjointe et $\mathbf{k}(x) = (x, \mathbf{A}(x))$ pour tous les x. Dans une base orthonormée, la matrice de la transformation **A** coïncide avec la matrice conjuguée de la matrice de **k**. Il s'ensuit que pour toute forme hermitienne **k** il existe une base orthonormée dans laquelle sa matrice est diagonale avec éléments diagonaux réels. En allongeant les vecteurs de base, on peut réduire la forme hermitienne à sa forme canonique, de sorte que les éléments diagonaux de la matrice associée deviennent égaux à 1, -1 ou 0. Les formes hermitiennes vérifient la loi d'inertie.

L'espace unitaire peut être défini comme un espace vectoriel complexe muni d'une forme hermitienne définie positive.

Pour un couple de formes hermitiennes dont l'une est définie positive on peut trouver une base dans laquelle elles se décomposent, toutes deux, en carrés.

CHAPITRE IX

ESPACES AFFINES

§ 1. Plans

1. Espace affine. Dans le chapitre I, on a admis que la notion d'espace géométrique ordinaire était familière au lecteur à partir de l'école secondaire et on a défini le vecteur comme un couple ordonné de points. Dans les chapitres VI à VIII, on a bâti la théorie des espaces vectoriels multidimensionnels. En partant de cette dernière, on est en mesure maintenant de fournir une définition axiomatique de l'espace ponctuel de dimension quelconque.

Considérons un espace vectoriel réel $\mathscr{L}_n$ de dimension n et introduisons la définition suivante.

DÉFINITION. On dit que l'ensemble $\mathscr{A}_n$ est l'*espace affine de dimension n* et ses éléments *points* s'il existe une relation qui à chaque couple ordonné de ses éléments A, B fait correspondre un vecteur de $\mathscr{L}_n$ et un seul (qu'on notera $\overrightarrow{AB}$) tel que :

1° Pour tout point A de $\mathscr{A}_n$ et tout vecteur x de $\mathscr{L}_n$ il existe un point B et un seul tel que $\overrightarrow{AB} = x$. Ce point sera noté $P(A, x)$.

2° Pour tout triplet de points A, B et C est vérifiée l'égalité $\overrightarrow{AB} + \overrightarrow{BC} = \overrightarrow{AC}$.

On dira que $\mathscr{L}_n$ est l'*espace de vecteurs* associé à l'espace $\mathscr{A}_n$, et ses éléments, les vecteurs dans $\mathscr{A}_n$.

Pour établir une correspondance avec les définitions du § 1, ch. I, remarquons que le premier axiome exprime la possibilité de construire un vecteur quelconque à partir d'un point arbitraire ; quant au second axiome, il répond à la définition de la somme des vecteurs.

Voici les plus simples corollaires :

a) Pour tout couple de points, $\overrightarrow{AA} + \overrightarrow{AB} = \overrightarrow{AB}$. Il s'ensuit que le vecteur associé à un couple de points confondus est nul. Pour tout point A on a $P(A, o) = A$, car $P(A, o) = P(A, \overrightarrow{AA}) = A$.

b) L'axiome 2° donne pour les points A, B, A la relation $\overrightarrow{AB} + \overrightarrow{BA} = \overrightarrow{AA}$, d'où $\overrightarrow{AB} = -\overrightarrow{BA}$.

c) Laissons au lecteur le soin de démontrer que pour quatre points A, B, A' et B', l'égalité $\overrightarrow{AB} = \overrightarrow{A'B'}$ est vérifiée si et seulement si $\overrightarrow{AA'} = \overrightarrow{BB'}$.

Cette propriété s'identifie à la définition de l'égalité des vecteurs, introduite au § 1, ch. I.

L'espace vectoriel $\mathscr{L}_n$ permet de construire un espace affine. En effet, soit $\mathscr{S}_n$ un ensemble de vecteurs dans l'espace $\mathscr{L}_n$. Associons à chaque couple de vecteurs x et y le vecteur $\overline{xy} = y - x$. On s'assure sans difficulté que les axiomes 1° et 2° se vérifient dans ce cas. Intuitivement cette construction nous suggère la représentation suivante. Si on interprète les vecteurs de $\mathscr{L}_n$ comme des segments orientés issus d'un même point, les extrémités de ces vecteurs sont des points de l'espace $\mathscr{S}_n$.

DÉFINITION. Les espaces affines $\mathscr{S}_n$ et $\mathscr{S}_n'$ sont dits *isomorphes* s'il existe une application bijective $\mathfrak{f} : \mathbb{S}_n \to \mathscr{S}_n'$ et un isomorphisme entre leurs espaces de vecteurs $\mathsf{F} : \mathscr{L}_n \to \mathscr{L}_n'$ tels que pour tous points A et B de $\mathscr{S}_n$ on a l'égalité $\overline{\mathfrak{f}(A)\mathfrak{f}(B)} = \mathsf{F}(\overline{AB})$.

Il ressort immédiatement de la définition que ne peuvent être isomorphes que des espaces de même dimension.

Si pour un point A on connaît son image $\mathfrak{f}(A)$ par l'isomorphisme $\mathfrak{f}$ et que soit donné l'isomorphisme F, l'application $\mathfrak{f}$ est définie de façon univoque. En effet, l'image de tout autre point B peut être obtenue par la formule $\mathfrak{f}(B) = P(\mathfrak{f}(A), \mathsf{F}(\overline{AB}))$. D'autre part, quels que soient $\mathfrak{f}(A)$ et F, on obtient nécessairement un isomorphisme $\mathfrak{f} : \mathscr{S}_n \to \mathscr{S}_n'$. En effet, on vérifie aisément que $\overline{\mathfrak{f}(A)\mathfrak{f}(B)} = \mathsf{F}(\overline{AB})$ et $\overline{\mathfrak{f}(A)\mathfrak{f}(C)} = \mathsf{F}(\overline{AC})$ entraînent $\overline{\mathfrak{f}(B)\mathfrak{f}(C)} = \mathsf{F}(\overline{BC})$ pour tous points B et C. Il en découle la

PROPOSITION 1. *Deux espaces affines de même dimension sont isomorphes. L'isomorphisme est défini univoquement si sont donnés l'image d'un des points et l'isomorphisme* $\mathsf{F} : \mathscr{L}_n \to \mathscr{L}_n'$.

Dans l'étude de chaque classe d'espaces, un rôle particulier est assumé par les transformations qui sont des isomorphismes appliquant ces espaces sur eux-mêmes. Pour les espaces vectoriels, ce sont des transformations linéaires bijectives ; pour les espaces euclidiens, des transformations orthogonales. Etudions les transformations isomorphes d'un espace affine.

A cet effet, supposons d'abord que l'isomorphisme $\mathsf{F} : \mathscr{L}_n \to \mathscr{L}_n$ est une transformation identique. Soit $\mathfrak{f}(A)$ l'image d'un point A. Considérons une transformation définie par l'égalité $\mathfrak{f}(B) = P(\mathfrak{f}(A), \overline{AB})$ quel que soit le point B. Notons pour abréger $\mathfrak{f}(A) = A^*$, $\mathfrak{f}(B) = B^*$. L'égalité précédente devient $\overline{A^*B^*} = \overline{AB}$. Or, elle est équivalente en vertu de c) à l'égalité $\overline{BB^*} = \overline{AA^*}$. Ainsi, l'image de chaque point s'obtient par *translation* de ce point de vecteur $\overline{AA^*}$.

Si l'on pose que $\mathfrak{f}(A) = A$ et F est une transformation linéaire bijective, on aboutit à une transformation de l'espace affine définie par la formule $\mathfrak{f}(B) = P(A, \mathsf{F}(\overline{AB}))$. Cette formule établit une correspondance biunivoque entre les transformations linéaires bijectives et les transformations isomorphes de l'espace affine laissant invariant le point A.

On voit qu'il existe, entre les espaces affines et vectoriels, une différence essentielle consistant dans le fait que l'ensemble des transformations isomorphes des premiers est plus étendu parce que comprend des translations.

DÉFINITION. L'espace affine est appelé *espace euclidien ponctuel* si son espace des vecteurs est euclidien.

Dans l'espace euclidien ponctuel, la longueur du vecteur $\overline{AB}$ s'appelle *distance entre les points A et B.*

16—6442

Un espace euclidien ponctuel tridimensionnel coïncide avec l'espace étudié en géométrie élémentaire si l'on y fixe l'unité de mesure.

On appelle *repère cartésien* de l'espace affine $\mathscr{S}_n$ un ensemble formé du point O et de la base $\boldsymbol{e}$ de l'espace de vecteurs $\mathscr{L}_n$. Si le repère $\{O, \boldsymbol{e}\}$ est donné, à chaque point A de l'espace $\mathscr{S}_n$ est associé de façon univoque un système ordonné de n nombres, à savoir, les composantes du vecteur $\overrightarrow{OA}$ dans la base $\boldsymbol{e}$. Ces nombres sont appelés *coordonnées cartésiennes* du point A dans le repère $\{O, \boldsymbol{e}\}$, et la matrice-colonne composée de ces nombres, *colonne de coordonnées* de A. Le point est défini de façon univoque par ses coordonnées si le repère est donné. Si $\boldsymbol{\xi}_1$ et $\boldsymbol{\xi}_2$ sont les colonnes de coordonnées des points A et B, la colonne des coordonnées du vecteur $\overrightarrow{AB}$ est $\boldsymbol{\xi}_2 - \boldsymbol{\xi}_1$. On le démontre comme au § 2 du ch. I.

La loi de transformation des coordonnées d'un point dans un changement de repère se déduit comme pour l'espace tridimensionnel au § 4 du ch. I.

2. Plans dans l'espace affine. Soient donnés le point A_0 de l'espace affine $\mathscr{S}_n$ et le sous-espace $\mathscr{L}_k$ de dimension k dans son espace de vecteurs $\mathscr{L}_n$. L'ensemble de tous les points $P(A_0, x)$, où x appartient à $\mathscr{L}_k$, est appelé *plan k-dimensionnel* dans $\mathscr{S}_n$. Il est évident que le point A_0 est aussi situé dans le plan. On l'appellera *point initial*, et le sous-espace $\mathscr{L}_k$, *sous-espace directeur*.

Tout point du plan $\mathscr{S}_k$ peut être pris pour son point initial. En effet, soit $A = P(A_0, x)$ un point de $\mathscr{S}_k$. Alors tout point $B = P(A_0, y)$ de $\mathscr{S}_k$ peut être représenté sous forme de $B = P(A, y - x)$, car $\overrightarrow{A_0B} - \overrightarrow{A_0A} = \overrightarrow{AB}$, $\overrightarrow{A_0B} = y$ et $\overrightarrow{A_0A} = x$, avec $y - x$ appartenant à $\mathscr{L}_k$. D'une façon analogue, pour chaque z de $\mathscr{L}_k$ on a $P(A, z) = P(A_0, z + x)$ et, par suite, $P(A, z)$ appartient au plan.

Il n'est pas difficile de démontrer que le plan de dimension k est un espace affine k-dimensionnel attaché à l'espace de vecteurs $\mathscr{L}_k$.

PROPOSITION 2. *Si on a choisi dans $\mathscr{S}_n$ un repère cartésien, les coordonnées des points de tout plan k-dimensionnel vérifient le système d'équations linéaires de rang $n - k$. Inversement, l'ensemble des points dont les coordonnées vérifient le système compatible d'équations linéaires de rang r est un plan $(n - r)$-dimensionnel.*

Pour le démontrer, désignons par $\boldsymbol{\xi}_0$ et $\boldsymbol{\xi}$ les matrices-colonnes des coordonnées respectives du point initial A_0 et d'un point variable A du plan $\mathscr{S}_k$. Puisque le vecteur $\overrightarrow{A_0A}$ appartient à $\mathscr{L}_k$, sa colonne de coordonnées $\boldsymbol{\xi} - \boldsymbol{\xi}_0$ vérifie le système homogène d'équations de rang $n - k$ (proposition 4, § 2, ch. VI). Soit U la matrice de ce système. Alors $U(\boldsymbol{\xi} - \boldsymbol{\xi}_0) = \boldsymbol{0}$ et $\boldsymbol{\xi}$ vérifie le système d'équations linéaires $U\boldsymbol{\xi} + \boldsymbol{\beta} = \boldsymbol{0}$, où $\boldsymbol{\beta} = -U\boldsymbol{\xi}_0$.

La seconde partie de la proposition découle du théorème 2, § 5, ch. V.

La solution générale du système d'équations linéaires (formule

(14), § 5, ch. V) nous fournit les équations paramétriques du plan $(n-r)$-dimensionnel dont les paramètres sont les coefficients de la combinaison linéaire du système fondamental de solutions, quant au système fondamental il joue le rôle de base dans le sous-espace directeur. Le point initial du plan est une solution particulière du système non homogène.

Le plan de dimension $n-1$ est appelé *hyperplan*. Il est défini par l'équation $\alpha_1 \xi^1 + \ldots + \alpha_n \xi^n + \beta = 0$.

Le plan unidimensionnel est appelé *droite*. Il est défini par le système d'équations de rang $n-1$ ou par l'équation paramétrique $\boldsymbol{\xi} = \boldsymbol{\xi}_0 + t\boldsymbol{\eta}$ dans laquelle $\boldsymbol{\xi}_0$ est la colonne de coordonnées du point initial et $\boldsymbol{\eta}$ la colonne de coordonnées d'un vecteur non nul du sous-espace directeur $\mathscr{L}_1$.

Etant donné un vecteur $\boldsymbol{\eta}$, l'ensemble de points de la droite qui correspondent à des valeurs $t \geqslant t_0$ est appelé *demi-droite*, et l'ensemble de points qui correspondent à $t_0 \leqslant t \leqslant t_1$, *segment*.

§ 2. Théorie générale des courbes et surfaces du deuxième ordre

On revient dans ce paragraphe à la géométrie de l'espace ponctuel tridimensionnel, à laquelle étaient consacrés les premiers chapitres du livre. L'étude de ce paragraphe peut donc être faite indépendamment du § 1. On y expose les applications des résultats obtenus pour les formes quadratiques dans les espaces vectoriels euclidiens à l'étude des courbes et surfaces du deuxième ordre.

1. Loi de transformation des coefficients. On commencera par des raisonnements applicables aussi bien aux courbes planes du deuxième ordre qu'aux surfaces du deuxième ordre dans l'espace. C'est pourquoi on ne précisera pas la dimension n qui peut être égale à deux ou trois suivant le cas considéré et on conviendra d'appeler les courbes et surfaces par un seul terme *surface* pour ne pas surcharger les énoncés.

Considérons une équation arbitraire du second degré

$$\sum_{i,j=1}^{n} \alpha_{ij} \xi^i \xi^j + 2 \sum_{i=1}^{n} \alpha_{i0} \xi^i + \alpha_{00} = 0, \tag{1}$$

reliant les coordonnées d'un point situé dans le plan ou dans l'espace. Ceci étant, aucune hypothèse ne sera faite ni sur ces points, ni sur leur existence. Si l'on change de repère et qu'on porte dans (1) l'expression des anciennes coordonnées du point courant en fonction des nouvelles, on obtient une nouvelle équation (également du second degré, en vertu des théorèmes 1 et 2 du § 1, ch. II). On dira dans ce cas que l'équation (1) a passé par changement des coordonnées à une nouvelle équation ou que ses coefficients se sont transformés.

Proposons-nous maintenant de rechercher une loi d'après laquelle se transforment les coefficients de l'équation du second degré lorsqu'on change de repère. Rappelons que le repère cartésien est composé d'un point (origine) et d'une base de l'espace vectoriel. Le changement de repère comprend donc une translation de l'origine et une transformation de la base (voir § 4 du ch. I).

Si l'on change de base, l'origine restant la même, les anciennes coordonnées s'expriment en fonction des nouvelles par la formule

$$\xi^i = \sum_{k=1}^{n} \sigma^i_k \xi'^k,$$

où σ^i_k sont les éléments de la matrice de passage de l'ancienne base à la nouvelle. En portant cette expression dans l'équation (1), on obtient l'équation du deuxième degré

$$\sum_{i,j,k,l} \alpha_{ij}\sigma^i_k\sigma^j_l\xi'^k\xi'^l + 2\sum_{i,k} \alpha_{i0}\sigma^i_k\xi'^k + \alpha_{00} = 0,$$

avec coefficients

$$\alpha'_{kl} = \sum_{i,j} \alpha_{ij}\sigma^i_k\sigma^j_l, \quad \alpha'_{k0} = \sum_i \alpha_{i0}\sigma^i_k, \quad \alpha'_{00} = \alpha_{00}. \tag{2}$$

Si l'on transporte l'origine des coordonnées en un point de coordonnées p^i $(1 \leqslant i \leqslant n)$ sans changer la base, les anciennes coordonnées s'expriment au moyen des nouvelles par la formule

$$\xi^i = \tilde{\xi}^i + p^i.$$

La substitution dans l'équation (1) donne

$$\sum_{i,j} \alpha_{ij}(\tilde{\xi}^i + p^i)(\tilde{\xi}^j + p^j) + 2\sum_i \alpha_{i0}(\tilde{\xi}^i + p^i) + \alpha_{00} = 0,$$

ou

$$\sum_{i,j} \alpha_{ij}\tilde{\xi}^i\tilde{\xi}^j + \sum_{i,j} \alpha_{ij}(\tilde{\xi}^i p^j + \tilde{\xi}^j p^i) + 2\sum_i \alpha_{i0}\tilde{\xi}^i + \tilde{\alpha}_{00} = 0.$$

D'où

$$\tilde{\alpha}_{ij} = \alpha_{ij}, \quad \tilde{\alpha}_{i0} = \sum_k \alpha_{ik}p^k + \alpha_{i0}, \tag{3}$$

vu que les sommes $\sum \alpha_{ij}\tilde{\xi}^i p^j$ et $\sum \alpha_{ij}\tilde{\xi}^j p^i$ ne diffèrent que par les notations des indices de sommation.

Les formules (2) et (3) expriment la loi de transformation des coefficients de l'équation (1) lorsqu'on change de repère. L'expression du terme constant $\tilde{\alpha}_{00}$ n'est pas nécessaire ici.

Les termes du second degré dans l'équation (1) constituent un polynôme homogène. On voit que ses coefficients α_{ij} ne varient pas par translation de l'origine des coordonnées, mais en revanche ils changent avec la base tout comme les coefficients de la forme quadratique. Aussi le polynôme

$$\sum_{i,j=1}^{n} \alpha_{ij}\xi^i\xi^j \tag{4}$$

peut-il être assimilé à une forme quadratique. Appelons-la *forme quadratique mineure.*

PROPOSITION 1. *Le rang et la signature de la forme quadratique mineure* (4) *ne varient pas dans tout changement de repère cartésien.*

Proposons-nous maintenant de déduire la loi de transformation des coefficients de l'équation (1) sous une autre forme pour pouvoir démontrer l'invariance de deux autres nombres.

Considérons un polynôme homogène du second degré de $n + 1$ variables

$$\sum_{p,q=0}^{n} \alpha_{pq}\xi^p\xi^q = \sum_{i,j=1}^{n} \alpha_{ij}\xi^i\xi^j + 2\sum_{i=1}^{n} \alpha_{i0}\xi^i\xi^0 + \alpha_{00}(\xi^0)^2. \tag{5}$$

Le premier membre de (1) résulte de (5) pour $\xi^0 = 1$.

Si l'on effectue une transformation linéaire biunivoque des variables, les coefficients du polynôme (5) se transforment comme les éléments de la matrice associée à la forme quadratique. Cette forme quadratique sera appelée *forme quadratique majeure.* De toutes les transformations on aura besoin de celles qui se présentent sous la forme (écrivons-la pour $n = 2$)

$$\left\|\begin{matrix}\xi^0\\ \xi^1\\ \xi^2\end{matrix}\right\| = \left\|\begin{matrix}1 & 0 & 0\\ \sigma_0^1 & \sigma_1^1 & \sigma_2^1\\ \sigma_0^2 & \sigma_1^2 & \sigma_2^2\end{matrix}\right\| \left\|\begin{matrix}\xi'^0\\ \xi'^1\\ \xi'^2\end{matrix}\right\|, \quad \begin{vmatrix}\sigma_1^1 & \sigma_2^1\\ \sigma_1^2 & \sigma_2^2\end{vmatrix} \neq 0. \tag{6}$$

La variable ξ^0 ne varie pas ici, mais les autres se transforment suivant les formules

$$\xi^i = \sum_{k=1}^{n} \sigma_k^i\xi'^k + \sigma_0^i\xi^0.$$

Si l'on pose $\xi^0 = 1$, on obtient la transformation la plus générale du repère cartésien. Ainsi, on a démontré la

PROPOSITION 2. *Le rang et la signature de la forme quadratique majeure (5) ne varient pas dans tout changement de repère cartésien.*

La surface définie par l'équation (1) ne varie pas si l'on multiplie l'équation par un facteur différent de zéro. Dans ce cas, les rangs des formes quadratiques (4) et (5) ne varient pas, quant aux signatures elles ne peuvent modifier que leur signe (si le facteur est strictement négatif). D'où le

THÉORÈME 1. *Les rangs et les modules des signatures des formes quadratiques majeure et mineure sont quatre invariants de la surface du deuxième ordre.*

On désignera le rang et le module de la signature de la forme quadratique (4) respectivement par r et σ, et le rang et le module de la signature de la forme (5) par R et Σ.

2. Courbes planes du deuxième ordre. Dans le chapitre III (théorème 1 du § 1) on a montré qu'une équation du second degré dans le plan, écrite sous forme générale, peut être réduite à l'une des neuf formes canoniques par le choix convenable d'un repère cartésien. Il existe par conséquent dans le plan sept classes de courbes du deuxième ordre (les équations de deux classes de courbes ne sont vérifiées par les coordonnées d'aucun point).

En composant les matrices associées aux formes quadratiques majeure et mineure pour les équations canoniques, on peut en déduire directement les valeurs de r, σ, R et Σ correspondant à chaque classe d'équations. L'unique difficulté se présente dans le cas de parabole. La matrice associée à sa forme quadratique majeure n'est pas diagonale et est de la forme

$$A = \begin{Vmatrix} 0 & 0 & -p \\ 0 & 1 & 0 \\ -p & 0 & 0 \end{Vmatrix}.$$

Pour trouver le rang R et le module de la signature Σ, transformons cette matrice suivant la formule ${}^tSAS$, avec

$$S = \begin{Vmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{Vmatrix}.$$

Il vient

$${}^tSAS = \begin{Vmatrix} 2p & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2p \end{Vmatrix}$$

et l'on découvre que $R = 3$ et $\Sigma = 1$. Le fait que la matrice S n'est pas de la

forme (6) est sans importance car R et Σ restent invariants dans toute transformation définie par une matrice de déterminant non nul.

Ecrivons les équations canoniques des courbes planes du deuxième ordre ainsi que les valeurs correspondantes des rangs et des modules de leurs signatures dans le tableau 1.

Tableau 1

Objet	Equation canonique	R	Σ	r	σ
Ellipse	$\frac{(\xi^1)^2}{a^2} + \frac{(\xi^2)^2}{b^2} = 1$	3	1	2	2
« Ellipse imaginaire »	$\frac{(\xi^1)^2}{a^2} + \frac{(\xi^2)^2}{b^2} = -1$	3	3	2	2
« Couple de droites sécantes imaginaires »	$a^2(\xi^1)^2 + b^2(\xi^2)^2 = 0$	2	2	2	2
Hyperbole	$\frac{(\xi^1)^2}{a^2} - \frac{(\xi^2)^2}{b^2} = 1$	3	1	2	0
Couple de droites sécantes	$a^2(\xi^1)^2 - b^2(\xi^2)^2 = 0$	2	0	2	0
Parabole	$(\xi^2)^2 = 2p\xi^1$	3	1	1	1
Couple de droites parallèles	$(\xi^2)^2 = a^2$	2	0	1	1
« Couple de droites parallèles imaginaires »	$(\xi^2)^2 = -a^2$	2	2	1	1
Deux droites confondues	$(\xi^2)^2 = 0$	1	1	1	1

On voit sur le tableau 1 qu'à toutes les équations d'une même classe correspondent les mêmes valeurs des invariants, tandis que les invariants correspondant aux équations des classes différentes sont distincts. Selon la proposition 8 du § 3, ch. IV, deux courbes quelconques d'une même classe peuvent être amenées l'une sur l'autre par une transformation affine convenablement choisie, mais aucune courbe ne peut devenir par transformation affine une courbe d'une autre classe. On obtient donc la

PROPOSITION 3. *Les nombres R, Σ, r et σ de deux courbes du deuxième ordre se confondent si et seulement si ces courbes peuvent être amenées l'une sur l'autre par une transformation affine.*

Si, comme au chapitre III, on n'utilise que les repères cartésiens rectangulaires, toute forme quadratique mineure peut être décomposée en carrés, sans que cette décomposition soit canonique. Par exemple, dans le cas de l'ellipse, l'équation canonique renferme les coefficients $1/a^2$ et $1/b^2$ affectant les carrés de ξ^1 et ξ^2. Cela est lié au fait que dans une base orthonormée la matrice de la forme quadratique doit coïncider avec la matrice de la transformation associée (comp. p. 235). Si donc on réduit la matrice

$$A = \begin{Vmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{12} & \alpha_{22} \end{Vmatrix} \tag{7}$$

de la forme quadratique (4) à une matrice diagonale par rapport à une base orthonormée, on doit avoir sur la diagonale les valeurs propres de la transformation associée.

Etant donné que dans tout changement de base orthonormée, la matrice A demeure attachée à une même transformation linéaire, les racines de son polynôme caractéristique ne varient pas lorsqu'on passe d'un repère cartésien rectangulaire à l'autre.

DÉFINITION. Les grandeurs ne variant pas lorsqu'on change de repère cartésien rectangulaire sont appelées *invariants orthogonaux* (ou *euclidiens*).

Au lieu des racines de l'équation caractéristique de la matrice (7) il est plus commode d'utiliser le système équivalent d'invariants orthogonaux, à savoir les coefficients de l'équation caractéristique de cette matrice :

$$I_1 = \alpha_{11} + \alpha_{22}, \quad I_2 = \begin{vmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{12} & \alpha_{22} \end{vmatrix}.$$

La transformation de variables (6) présente une forme spéciale, de sorte que la forme quadratique majeure ne peut en général être décomposée en carrés (cas de $R = 3, r = 1$, c'est-à-dire de la parabole). Toutefois, si le repère rectangulaire se transforme en un repère rectangulaire, la matrice

$$\begin{Vmatrix} \sigma_1^1 & \sigma_2^1 \\ \sigma_1^2 & \sigma_2^2 \end{Vmatrix} \tag{8}$$

est une matrice orthogonale et son déterminant vaut 1 ou -1. Dans ce cas, le déterminant de la matrice de passage de (6) est également ± 1 et par suite, le déterminant de la matrice

$$\tilde{A} = \begin{Vmatrix} \alpha_{00} & \alpha_{10} & \alpha_{20} \\ \alpha_{10} & \alpha_{11} & \alpha_{12} \\ \alpha_{20} & \alpha_{12} & \alpha_{22} \end{Vmatrix} \tag{9}$$

associée à la forme quadratique (5) demeure invariant par transformation (6). On a obtenu encore un invariant orthogonal de l'équation du deuxième degré, soit :

$$I_3 = \det \tilde{A}.$$

On constate sans difficulté que la matrice de passage dans la formule (6) est orthogonale si et seulement si est orthogonale la matrice (8) et $\sigma_0^1 = \sigma_0^2 = 0$, autrement dit, si la base orthonormée est remplacée par une base orthonormée et l'origine des coordonnées n'est pas déplacée. Les coefficients du polynôme caractéristique de la matrice $\tilde{A}$ demeurent invariants si la matrice de passage est orthogonale, car dans ce cas $\tilde{A}$ reste égale à la matrice de la transformation associée à la forme (5). Donc les grandeurs

$$\alpha_{11} + \alpha_{22} + \alpha_{00} \tag{10}$$

et

$$\begin{vmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{12} & \alpha_{22} \end{vmatrix} + \begin{vmatrix} \alpha_{00} & \alpha_{10} \\ \alpha_{10} & \alpha_{11} \end{vmatrix} + \begin{vmatrix} \alpha_{00} & \alpha_{20} \\ \alpha_{20} & \alpha_{22} \end{vmatrix} \tag{11}$$

qui sont respectivement égales aux coefficients de λ^2 et de $-\lambda$ dans l'équation caractéristique ne varient pas lorsqu'on passe d'une base orthogonale à l'autre, et, probablement, varient avec le déplacement de l'origine des coordonnées. Les grandeurs de ce genre sont dites *semi-invariantes*. En retranchant les invariants I_1 et I_2 de (10) et (11) respectivement, on obtient les

semi-invariants α_{00} et

$$K_1 = \begin{vmatrix} \alpha_{00} & \alpha_{10} \\ \alpha_{10} & \alpha_{11} \end{vmatrix} + \begin{vmatrix} \alpha_{00} & \alpha_{20} \\ \alpha_{20} & \alpha_{22} \end{vmatrix}.$$

D'ailleurs, le fait que α_{00} est un semi-invariant se déduit des formules (2).

Les valeurs obtenues des invariants orthogonaux et des semi-invariants permettent de trouver les coefficients a, b et p dans les équations canoniques et définissent ainsi une courbe du deuxième ordre à sa position dans le plan près (voir propositon 9, § 3, ch. IV). Le calcul des coefficients d'équations canoniques d'après les invariants orthogonaux est effectué dans tous les cours détaillés de géométrie analytique.

Il ne faut pas oublier que les grandeurs I_1, I_2, I_3 et K_1 sont liées au polynôme du deuxième degré et non pas à la courbe. Elles varient de façon évidente si on multiplie l'équation par un nombre différent de zéro.

3. Surfaces du deuxième ordre. Admettons que l'équation (1) relie les coordonnées d'un point dans un espace tridimensionnel. On montrera qu'il existe un repère cartésien rectangulaire dans lequel l'équation prend l'une des 17 formes canoniques.

Choisissons pour base du repère recherché une base orthonormée dans laquelle la forme quadratique mineure se décompose en carrés. Ainsi donc, on partira de l'équation

$$\lambda_1(\xi^1)^2 + \lambda_2(\xi^2)^2 + \lambda_3(\xi^3)^2 + 2\alpha_{10}\xi^1 + 2\alpha_{20}\xi^2 + 2\alpha_{30}\xi^3 + \alpha_{00} = 0 \quad (12)$$

sans oublier que la base orthonormée est choisie. Aucune contrainte n'est imposée aux coefficients, à l'exception de ce que λ_1, λ_2 et λ_3 ne s'annulent pas en même temps. Les simplifications ultérieures se définissent par la proposition auxiliaire suivante.

PROPOSITION 4. *Si l'équation* (12) *contient le carré* (*avec un coefficient non nul*) *d'une coordonnée, on peut, par déplacement de l'origine des coordonnées le long de l'axe correspondant, rendre nul le terme renfermant la puissance première de cette coordonnée.*

Cette proposition se démontre comme la proposition 1 du § 1, ch. III.

Considérons pour plus de commodité quelques cas qui correspondent aux différentes valeurs des invariants R, Σ, r et σ.

1. *Posons* $r = 3$. Alors, en vertu de la proposition 4, l'origine des coordonnées peut être transportée en un point pour lequel l'équation (12) se met sous la forme

$$\lambda_1(\xi^1)^2 + \lambda_2(\xi^2)^2 + \lambda_3(\xi^3)^2 + \mu = 0, \quad (13)$$

où λ_1, λ_2 et λ_3 ne sont pas nuls.

1A. La condition $R = 4$ s'identifie au cas où le terme constant μ de (13) est différent de zéro. En divisant par ce terme, on obtient

$$-\frac{\lambda_1}{\mu}(\xi^1)^2 - \frac{\lambda_2}{\mu}(\xi^2)^2 - \frac{\lambda_3}{\mu}(\xi^3)^2 = 1. \quad (14)$$

1Aa. Posons $\Sigma = 4$. Il vient alors que λ_1, λ_2, λ_3 et μ sont de même signe, de sorte que les coefficients de (14) sont strictement négatifs. L'équation se réduit à la forme canonique

$$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} + \frac{(\xi^3)^2}{\gamma^2} = -1.$$

Elle n'est vérifiée par aucun point. Cette équation est l'équation de l'*ellipsoïde imaginaire*.

1Ab. Si $\Sigma = 2$ et $\sigma = 3$, le signe commun de λ_1, λ_2 et λ_3 est opposé à celui de μ. Les coefficients dans (14) sont positifs. L'équation se réduit à la forme canonique

$$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} + \frac{(\xi^3)^2}{\gamma^2} = 1.$$

La surface est un *ellipsoïde*.

1Ac. Si $\Sigma = 0$ et $\sigma = 1$, le signe d'une des valeurs propres (sans restreindre la généralité on peut poser que c'est λ_3) est opposé au signe commun des deux autres (λ_1 et λ_2) et coïncide avec celui de μ. Dans l'équation (14) il y a deux coefficients positifs et un négatif. La forme canonique de l'équation est

$$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} - \frac{(\xi^3)^2}{\gamma^2} = 1.$$

La surface est un *hyperboloïde à une nappe*.

1Ad. Soit maintenant $\Sigma = 2$, $\sigma = 1$. Le signe de l'une des valeurs propres (par exemple, de λ_1) est opposé au signe commun des deux autres (λ_2 et λ_3) et est opposé au signe de μ. Dans l'équation (14) il y a deux coefficients négatifs et un positif. Elle se réduit à la forme

$$\frac{(\xi^1)^2}{\alpha^2} - \frac{(\xi^2)^2}{\beta^2} - \frac{(\xi^3)^2}{\gamma^2} = 1$$

et définit un *hyperboloïde à deux nappes*.

1B. Posons $R = 3$. Dans le cas considéré on a $r = 3$, ce qui équivaut à $\mu = 0$; l'équation (13) pour $R = 3$ est homogène. Signalons qu'on a obligatoirement $\Sigma = \sigma$.

1Ba. Pour $\sigma = 3$, toutes les valeurs propres λ_1, λ_2 et λ_3 dans l'équation (13) ont le même signe, et l'équation peut être écrite sous la forme

$$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} + \frac{(\xi^3)^2}{\gamma^2} = 0.$$

On la dénomme équation du *cône imaginaire*. La surface est réduite à un seul point.

1Bb. Si $\sigma = 1$, une des valeurs propres diffère de signe des deux autres. L'équation se réduit à la forme canonique suivante

$$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} - \frac{(\xi^3)^2}{\gamma^2} = 0.$$

La surface est un *cône du deuxième ordre*.

2. *Posons maintenant* $r = 2$. Dans l'équation (12), l'une des valeurs propres est nulle. Sans restreindre la généralité on peut admettre que c'est λ_3. En recourant à la proposition 4, réduisons l'équation à la forme

$$\lambda_1(\xi'^1)^2 + \lambda_2(\xi'^2)^2 + 2\alpha_{30}\xi'^3 + \alpha'_{00} = 0. \tag{15}$$

On transporte dans ce cas l'origine des coordonnées le long des axes ξ^1 et ξ^2. Ecrivons le déterminant de la matrice de la forme quadratique (5) pour l'équation (15) :

$$\begin{vmatrix} \alpha'_{00} & 0 & 0 & \alpha_{30} \\ 0 & \lambda_1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 0 \\ \alpha_{30} & 0 & 0 & 0 \end{vmatrix} = -\alpha_{30}^2\lambda_1\lambda_2. \tag{16}$$

2A. La condition $R = 4$ équivaut à $\alpha_{30} \neq 0$ en vertu de l'égalité (16). On peut mettre l'équation (15) sous la forme

$$\lambda_1(\xi'^1)^2 + \lambda_2(\xi'^2)^2 + 2\alpha_{30}(\xi'^3 + \alpha'_{00}/2\alpha_{30}) = 0,$$

ce qui montre que la translation de l'origine des coordonnées le long de l'axe ξ'^3 définie par

$$\tilde{\xi}^1 = \xi'^1, \tilde{\xi}^2 = \xi'^2, \quad \tilde{\xi}^3 = \xi'^3 + \alpha'_{00}/2\alpha_{30},$$

permet de transformer l'équation en

$$\lambda_1(\tilde{\xi}^1)^2 + \lambda_2(\tilde{\xi}^2)^2 + 2\alpha_{30}\tilde{\xi}^3 = 0. \tag{17}$$

Ensuite, deux possibilités peuvent se présenter suivant les valeurs de l'invariant σ.

2Aa. $\sigma = 2$. Dans ce cas, λ_1 et λ_2 sont de même signe. En substituant, si c'est nécessaire, le vecteur de base $-\mathbf{e}_3$ à $\mathbf{e}_3$, on réduit l'équation à la forme

$$\frac{(\tilde{\xi}^1)^2}{\alpha^2} + \frac{(\tilde{\xi}^2)^2}{\beta^2} = 2\tilde{\xi}^3.$$

C'est l'équation canonique du *paraboloïde elliptique*.

2Ab. $\sigma = 0$. Dans ce cas, λ_1 et λ_2 sont de signes différents, et l'équation se réduit à la forme canonique suivante

$$\frac{(\tilde{\xi}^1)^2}{\alpha^2} - \frac{(\tilde{\xi}^2)^2}{\beta^2} = 2\tilde{\xi}^3$$

(ici de même, il se peut qu'il faille modifier l'orientation de $\mathbf{e}_3$). Cette équation définit un *paraboloïde hyperbolique*.

2B. Supposons maintenant que $R = 3$. Il ressort alors de (16) que pour ces équations on a $\alpha_{30} = 0$, de sorte que le premier membre ne contient pas la variable ξ^3. En accord avec ce qui a été dit au point 5 du § 1, ch. II, cela signifie que l'équation définit un *cylindre* dont les génératrices sont parallèles au vecteur de base $\mathbf{e}_3$ et la directrice est définie dans le plan des vecteurs $\mathbf{e}_1$ et $\mathbf{e}_2$ par l'équation

$$\lambda_1(\xi'^1)^2 + \lambda_2(\xi'^2)^2 + \alpha'_{00} = 0, \tag{18}$$

qui est l'équation (15) pour $\alpha_{30} = 0$.

L'équation (18) définit dans le plan une des cinq coniques à centre. Il leur correspond cinq cylindres que cette équation définit dans l'espace : *cylindre elliptique, cylindre hyperbolique, couple de plans sécants* (la directrice est ici le couple de droites sécantes), *couple de plans sécants imaginaires* (la surface se réduit à une droite et la directrice à un point, autrement dit à un « couple de droites sécantes imaginaires ») et, enfin « *cylindre elliptique imaginaire* » (la surface ne comprend aucun point et la « directrice » est une ellipse imaginaire). Les équations canoniques sont données dans le tableau 2.

3. *Considérons maintenant le cas de* $r = 1$. Choisissons une base orthonormée de manière que la forme quadratique (4) se décompose en carrés. On aboutit à l'équation

$$\lambda_1(\xi^1)^2 + 2\alpha_{10}\xi^1 + 2\alpha_{20}\xi^2 + 2\alpha_{30}\xi^3 + \alpha_{00} = 0, \tag{19}$$

où λ_1 est différent de zéro.

3A. Supposons que $\alpha_{20}^2 + \alpha_{30}^2 \neq 0$. On peut alors faire subir à la base une rotation autour de $\mathbf{e}_1$:

$$\xi'^1 = \xi^1, \quad \xi'^2 = (\alpha_{20}\xi^2 + \alpha_{30}\xi^3)/v, \quad \xi'^3 = (-\alpha_{30}\xi^2 + \alpha_{20}\xi^3)/v,$$

où $v = \sqrt{\alpha_{20}^2 + \alpha_{30}^2}$. L'équation (19) prend alors la forme

$$\lambda_1(\xi'^1)^2 + 2\alpha_{10}\xi'^1 + 2v\xi'^2 + \alpha_{00} = 0,$$

qui, en vertu de la proposition 4, se transforme en

$$\lambda_1(\xi''^1)^2 + 2v\xi'^2 + \alpha'_{00} = 0.$$

Ensuite, la translation de l'origine le long de l'axe ξ^2, définie par

$$\tilde{\xi}^1 = \xi''^1, \qquad \tilde{\xi}^2 = \xi'^2 + \alpha'_{00}/2v, \qquad \tilde{\xi}^3 = \xi^3,$$

transforme l'équation en

$$\lambda_1(\tilde{\xi}^1)^2 + 2v\tilde{\xi}^2 = 0. \tag{20}$$

L'équation (20) se réduit à la forme canonique $(\tilde{\xi}^1)^2 = 2\alpha\tilde{\xi}^2$, où $\alpha > 0$. (Si

nécessaire, on peut modifier l'orientation de e'_2.) C'est l'équation du *cylindre parabolique*.

3B. Si $\alpha_{20} = \alpha_{30} = 0$, l'équation (19) ne contient que ξ^1 et se réduit facilement à l'une des trois dernières formes canoniques données dans le tableau 2.

On achève ainsi la répartition en classes dont les résultats sont groupés dans le tableau 2.

Tableau 2

Objet	Equation canonique	R	Σ	r	σ
« Ellipsoïde imaginaire »	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} + \frac{(\xi^3)^2}{\gamma^2} = -1$	4	4	3	3
Ellipsoïde	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} + \frac{(\xi^3)^2}{\gamma^2} = 1$	4	2	3	3
Hyperboloïde à une nappe	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} - \frac{(\xi^3)^2}{\gamma^2} = 1$	4	0	3	1
Hyperboloïde à deux nappes	$\frac{(\xi^1)^2}{\alpha^2} - \frac{(\xi^2)^2}{\beta^2} - \frac{(\xi^3)^2}{\gamma^2} = 1$	4	2	3	1
« Cône imaginaire »	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} + \frac{(\xi^3)^2}{\gamma^2} = 0$	3	3	3	3
Cône	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} - \frac{(\xi^3)^2}{\gamma^2} = 0$	3	1	3	1
Paraboloïde elliptique	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} = 2\xi^3$	4	2	2	2
Paraboloïde hyperbolique	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} = 2\xi^3$	4	0	2	0
Cylindre elliptique	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} = 1$	3	1	2	2
« Cylindre elliptique imaginaire »	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} = -1$	3	3	2	2
Cylindre hyperbolique	$\frac{(\xi^1)^2}{\alpha^2} - \frac{(\xi^2)^2}{\beta^2} = 1$	3	1	2	0
Couple de plans sécants	$\frac{(\xi^1)^2}{\alpha^2} - \frac{(\xi^2)^2}{\beta^2} = 0$	2	0	2	0
« Couple de plans sécants imaginaires »	$\frac{(\xi^1)^2}{\alpha^2} + \frac{(\xi^2)^2}{\beta^2} = 0$	2	2	2	2
Cylindre parabolique	$(\xi^1)^2 = 2\alpha\xi^2$	3	1	1	1
Couple de plans parallèles	$(\xi^1)^2 - \alpha^2 = 0$	2	0	1	1
« Couple de plans parallèles imaginaires »	$(\xi^1)^2 + \alpha^2 = 0$	2	2	1	1
Couple de plans confondus	$(\xi^1)^2 = 0$	1	1	1	1

CHAPITRE X

ÉLÉMENTS D'ALGÈBRE TENSORIELLE

§ 1. Tenseurs dans l'espace vectoriel

1. Objets géométriques. Dans les chapitres VI à VIII on a étudié les espaces vectoriels et les objets de nature différente définis dans ces espaces : transformations linéaires, formes quadratiques, etc. L'étude de chaque objet s'appuyait sur une définition formulée indépendamment de la base dans l'espace. Par exemple, la transformation linéaire était définie comme une application de l'espace dans lui-même qui satisfait à deux conditions (comp. (1), § 3, ch. VI). Ainsi donc, ces objets existent et, en principe, peuvent être étudiés comme les objets en géométrie sans qu'on introduise une base. On dira que ce sont des *objets géométriques*.

Bien qu'un objet géométrique existe indépendamment de la base, il est utile de choisir une base et de définir cet objet par rapport à cette base au moyen d'un système ordonné de nombres appelés *composantes* de l'objet. Par exemple, le choix d'une base établit une correspondance biunivoque entre les transformations linéaires et les matrices carrées. Une transformation linéaire d'un espace n-dimensionnel possède n^2 composantes. Le fait qu'un objet reste invariant dans tout changement de base entraîne la variation de ses composantes. Dans tous les cas rencontrés, on a pu calculer les composantes de l'objet dans une base en fonction de ses composantes dans une autre base et des éléments de la matrice de passage d'une base à l'autre.

Outre les objets géométriques étudiés, il en existe beaucoup d'autres. Il n'est pas toujours commode d'en donner une définition indépendante de la base. Parfois un objet géométrique est défini au moyen de ses composantes. Les composantes de l'objet géométrique diffèrent d'un ensemble arbitraire de nombres par la propriété suivante qu'on prend pour définition rigoureuse.

DÉFINITION. On dira que dans un espace vectoriel est défini un *objet géométrique* si à toute base correspond une famille ordonnée de nombres (composantes) et une seule et si les composantes dans une base peuvent être exprimées en fonction des composantes dans une autre base et des éléments de la matrice de passage. Cette relation entre les composantes de deux bases est appelée *loi de transformation des composantes* de l'objet géométrique.

Si à toute base, par exemple, correspond un même nombre, on dit qu'est donné un objet géométrique à une seule composante, appelé *invariant*. La loi de transformation se traduit ici par le fait que cette composante ne varie pas dans tout changement de base.

A titre d'un autre exemple, considérons la forme quadratique **k** et faisons correspondre à chaque base e le déterminant δ de sa matrice K dans cette base. On obtient un objet géométrique à une seule composante $\delta =$ $= \det K$. Vu que dans la base $e' = eS$ on a $K' = {}^tSKS$, la loi de transformation de la composante de cet objet est de la forme $\delta' = \delta(\det S)^2$.

Soit donné le vecteur x. Associons à la base e la somme des composantes ξ^i de x dans cette base. A chaque base sera associé un nombre bien déterminé $\gamma = \sum\limits_i \xi^i$. Cependant, une telle correspondance ne définit pas un objet géométrique. En effet, le nombre γ' correspondant à la base e' s'exprime non pas en fonction de γ mais en fonction des composantes du vecteur :

$$\gamma' = \sum_i \xi'^i = \sum_{i,j} \tau^i_j \xi^j.$$

(τ^i_j sont ici les éléments de la matrice de passage de e' à e.)

Dans ce chapitre, on étudiera un espace vectoriel réel n-dimensionnel $\mathscr{L}_n$. On conservera toutes les notations adoptées auparavant. Si l'on a deux bases e et e', la matrice de passage sera toujours notée S, et ses éléments, σ^i_j, de sorte que $e' = eS$ ou $e'_j = \sum\limits_i \sigma^i_j e_i$. Les éléments de la matrice de passage inverse S^{-1} seront notés τ^i_j.

Considérons un objet géométrique quelconque. Désignons ses composantes dans la base e par $\alpha_1, \ldots, \alpha_N$ et les composantes dans la base e' par $\alpha'_1, \ldots, \alpha'_N$. La loi de transformation des composantes se définit par N fonctions $F_1, \ldots, F_N$ dont chacune dépend de $N + n^2$ variables indépendantes :

$$\alpha'_K = F_K(\alpha_1, \ldots, \alpha_N, \sigma^1_1, \ldots, \sigma^n_n) \quad (K = 1, \ldots, N),$$

ou d'une façon plus concise

$$\alpha'_K = F_K(\alpha, S). \tag{1}$$

Considérons une troisième base $e'' = e'R = eSR$. Soient $\alpha''_1, \ldots, \alpha''_N$ les composantes de l'objet dans cette base. On peut passer de e à e'' directement, la matrice de passage étant le produit SR. On obtient

$$\alpha''_K = F_K(\alpha, SR).$$

D'autre part, on peut passer de e à e' et de e' à e''. Il vient alors

$$\alpha''_K = F_K(\alpha', R) = F_K(F_1(\alpha, S), \ldots, F_N(\alpha, S), R).$$

Vu que les composantes de l'objet se définissent par la base de façon univoque, les fonctions F_K pour tous $K = 1, \ldots, N$ doivent satisfaire à la condition

$$F_K(\alpha, SR) = F_K(F_1(\alpha, S), \ldots, F_N(\alpha, S), R), \qquad (2)$$

quels que soient α, S et R. On a obtenu la

PROPOSITION 1. *Si les fonctions de la forme* (1) *déterminent la loi de transformation d'un objet géométrique, elles doivent satisfaire à la condition* (2).

L'assertion réciproque est également vraie.

PROPOSITION 2. *Toute famille de fonctions de la forme* (1) *satisfaisant à la condition* (2) *définit la loi de transformation d'un objet géométrique.*

Pour le démontrer, construisons cet objet géométrique de la façon suivante. Soient des composantes arbitraires par rapport à une base e (on suppose que les fonctions considérées sont définies pour toutes les valeurs des variables indépendantes). Cherchons les composantes de l'objet par rapport à d'autres bases en traitant la famille des fonctions données comme loi de transformation. Dans ce cas, étant donné deux bases quelconques e' et e'', les composantes dans e'' se déterminent par les mêmes fonctions des composantes dans e' et par les éléments de la matrice de passage de e' à e''. En effet, en utilisant les notations introduites plus haut, on a $\alpha''_K = F_K(\alpha, SR)$ et $\alpha'_K = F_K(\alpha, S)$. D'où, en vertu de (2), $\alpha''_K = F_K(\alpha', R)$.

Il reste à montrer que la transformation ainsi définie fait associer à chaque base une seule famille de nombres. L'unicité ne peut être perturbée que par le fait qu'il existe plusieurs façons pour obtenir une même base : étant donné des bases quelconques $e', e'', \ldots, e^{(k)}$, on peut passer de e à e', de e' à e'', etc. et, enfin, de $e^{(k)}$ à f. En appliquant plusieurs fois la condition (2), on se convainc que tout passage à la base f fournit les mêmes composantes par rapport à f que le passage direct de e à f.

On étudiera les objets géométriques les plus simples appelés *tenseurs*. La loi de transformation de leurs composantes est telle que les nouvelles composantes sont des polynômes homogènes linéaires d'anciennes composantes. On la décrira en détail plus loin. On a pris l'habitude d'ordonner les composantes des tenseurs en les numérotant avec plusieurs indices. Le point suivant sera consacré aux notations.

2. Matrices multidimensionnelles. Rappelons qu'à la page 122 on a défini la matrice comme fonction associant un nombre à chaque couple ordonné (i, j), où $i \in \{1, \ldots, m\}$, $j \in \{1, \ldots, n\}$. On généralisera cette définition. Puisqu'on n'aura besoin que de matrices analogues aux matrices carrées, tous les indices appartiendront à un même ensemble $\{1, \ldots, n\}$.

DÉFINITION. On appelle *matrice s-dimensionnelle d'ordre n* la fonction sur l'ensemble des s-uples de la forme $(i_1, \ldots, i_s)$, où $i_1, \ldots, i_s$ prennent les valeurs $1, \ldots, n$.

Soulignons que l'ordre de la matrice coïncide avec la dimension de l'espace vectoriel considéré. Pour expliquer l'origine du terme « matrice s-dimensionnelle », considérons la matrice tridimensionnelle d'éléments a_{ijk}. Pour toute valeur fixée de l'indice $k = k_0$, les éléments de type a_{ijk_0} constituent une matrice carrée d'ordre n. Ainsi, tous les éléments a_{ijk} se répartissent en n matrices carrées d'ordre n : $\| a_{ij1} \|, \ldots, \| a_{ijn} \|$. On peut se

représenter ces matrices comme des couches disposées l'une au-dessous l'autre, en formant de la sorte un cube qui contient n^3 blocs dans chacun desquels est inscrit un nombre. D'une façon analogue, une matrice quadridimensionnelle peut être considérée comme un n-uple ordonné de n matrices tridimensionnelles, etc.

Il est commode d'assimiler les matrices-lignes et les matrices-colonnes à des matrices unidimensionnelles. Leurs éléments sont numérotés avec un seul indice.

En choisissant les notations pour la matrice s-dimensionnelle, on peut pour plus de commodité, convenir d'écrire certains indices en haut et les autres en bas. Mais, une fois la notation adoptée, on observera rigoureusement la disposition choisie pour les indices. Si l'ordre des indices n'est pas réglementé, on considérera que les indices inférieurs suivent les indices supérieurs comme s'ils étaient écrits à droite de ces derniers.

Les matrices multidimensionnelles sont difficiles à écrire d'une façon explicite. On a convenu de considérer tout indice désigné par une lettre comme une variable prenant les valeurs 1, ..., n. *Si est écrite une expression contenant un indice littéral* *) *qui n'est pas un indice de sommation, on admet que sont écrites n expressions semblables pour chaque valeur de cet indice.* S'il y a plusieurs indices, la convention ci-dessus se rapporte à chacun d'entre eux. Par exemple, $\alpha^{i_1 \dots i_s}$ désigne l'ensemble de tous les éléments de la matrice s-dimensionnelle, et la relation $\alpha^i_{jk} = \beta^i_{jk}$ veut dire que sont égaux les éléments qui occupent la même place dans deux matrices tridimensionnelles, ou bien que les matrices sont égales.

Introduisons une nouvelle notation de sommation. *Soit une expression contenant une lettre ou un produit de plusieurs lettres affectées d'indices dont l'un figure deux fois : une fois en haut et une fois en bas. On entendra par cette expression la somme des termes de cette forme, écrits pour toutes les valeurs de l'indice qui se répète, sans écrire le symbole de sommation.* D'une façon analogue, s'il y a plusieurs indices qui se répètent, on a en vue une somme multiple. Dans les chapitres V à VIII, on a constamment rencontré des sommes de ce genre mais on écrivait le symbole de sommation. Dorénavant, on ne le fera plus. Par exemple, les formules

$$\mathbf{f}(x) = \sum_{i=1}^{n} x_i \xi^i, \quad \alpha'_{kl} = \sum_{i,j} \alpha_{ij} \sigma^i_k \sigma^j_l$$

seront écrites sous la forme

$$\mathbf{f}(x) = x_i \xi^i, \quad \alpha'_{kl} = \alpha_{ij} \sigma^i_k \sigma^j_l. \tag{3}$$

*) Pour désigner les indices, nous utiliserons le plus souvent les lettres i, j, k, l, affectées quelquefois de leurs propres indices. La lettre n désignera toujours un nombre fixé, la dimension de l'espace.

17—6442

3. Définition et exemples. Soit un espace vectoriel réel $\mathscr{L}_n$ de dimension n.

DÉFINITION. On dit qu'un *tenseur de type* (p, q) est défini dans $\mathscr{L}_n$ si à chaque base est associé une matrice $(p + q)$-dimensionnelle d'ordre n. Ceci étant, quelles que soient les bases e et e', les éléments $\alpha_{j_1 \ldots j_q}^{i_1 \ldots i_p}$ et $\alpha_{j_1 \ldots j_q}^{' i_1 \ldots i_p}$ des matrices associées respectivement à e et e' doivent être liés par les relations

$$\alpha_{j_1 \ldots j_q}^{' i_1 \ldots i_p} = \tau_{k_1}^{i_1} \ldots \tau_{k_p}^{i_p} \sigma_{j_1}^{l_1} \ldots \sigma_{j_q}^{l_q} \alpha_{l_1 \ldots l_q}^{k_1 \ldots k_p}, \tag{4}$$

où σ_j^i sont les éléments de la matrice de passage de e à e', et τ_k^i les éléments de sa matrice inverse.

Dans le second membre de (4), la sommation est réalisée suivant tous les indices k et l, autrement dit on a une somme de multiplicité $p + q$. On conviendra d'écrire en haut les indices auxquels, dans chaqun des n^{p+q} termes, correspondent les facteurs τ_k^i. Ces indices sont dénommés *contravariants*. Par contre, on écrira en bas les indices auxquels, dans chaque terme, correspondent les facteurs σ_j^l. Ce sont les indices *covariants*.

Les éléments de la matrice associée à une base dans $\mathscr{L}_n$ s'appellent *composantes* du tenseur dans cette base.

On dit que le nombre $p + q$ est *valence* du tenseur, et que q et p sont respectivement *valence covariante* et *valence contravariante*.

Soulignons que malgré la complexité de la somme dans le second membre de (4), chaque terme contient une seule composante ancienne du tenseur. Cela veut dire que les nouvelles composantes s'expriment en fonction des anciennes par des polynômes homogènes linéaires. La complexité de la formule (4) est due à l'expression des coefficients de ces polynômes par l'intermédiaire d'éléments de la matrice de passage.

Deux tenseurs sont *égaux* s'ils sont de même type et possèdent des composantes égales dans une certaine base. Il découle alors de la loi de transformation que leurs composantes sont égales dans toute base.

EXEMPLE 1. Le vecteur est un tenseur de type (1, 0). En effet, étant donné un vecteur, à chaque base correspond une matrice unidimensionnelle, la matrice-colonne. Ceci étant, les éléments des matrices-colonnes associées à des bases égales sont liés par la formule (4) du § 1, ch. VI :

$$\xi^i = \sigma_k^i \xi'^k.$$

Cherchons, en partant de cette relation, l'expression de ξ'^k en fonction de ξ^i :

$$\xi'^k = \tau_i^k \xi^i,$$

c'est la loi de transformation des composantes du tenseur de type (1, 0).

EXEMPLE 2. La fonction linéaire sur l'espace $\mathscr{L}_n$ est un tenseur de type (0, 1). En effet, étant donné une fonction linéaire, à chaque base correspond une matrice unidimensionnelle, la matrice-ligne des coefficients de cette fonction. Lorsqu'on change de base, les coefficients de la fonction linéaire se transforment suivant la formule (4) du § 1, ch. VIII :

$$x_i' = \sigma_i^k x_k.$$

C'est justement la loi de transformation des composantes du tenseur de type (0, 1). Les tenseurs de type (0, 1) sont appelés *covecteurs*.

EXEMPLE 3. La transformation linéaire de l'espace $\mathscr{L}_n$ est un tenseur de type (1, 1). En effet, étant donné une transformation linéaire, à chaque base correspond une matrice bidimensionnelle d'ordre n. Si la base change, les éléments de la matrice se transforment suivant la formule (1) du § 4, ch. VI : $A' = S^{-1}AS$, ou

$$\alpha_j'^i = \tau_k^i \sigma_j^l \alpha_l^k.$$

EXEMPLE 4. La forme bilinéaire sur l'espace $\mathscr{L}_n$ est un tenseur de type (0, 2). En effet, étant donné une forme bilinéaire, à chaque base correspond une matrice bidimensionnelle d'ordre n. Les éléments des matrices associées à deux bases sont liés par la formule :

$$\beta_{ij}' = \sigma_i^k \sigma_j^l \beta_{kl}.$$

Signalons que la forme bilinéaire symétrique et la forme quadratique qui lui correspond représentent le même tenseur, vu que leurs matrices se confondent dans toute base.

EXEMPLE 5. Un invariant peut être assimilé à un tenseur de type (0, 0).

EXEMPLE 6. Un tenseur important de type (1, 1) est le *symbole de Kronecker* dont les composantes constituent, dans une certaine base, la matrice unité :

$$\delta_j^i = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases} \tag{5}$$

Selon la loi de transformation, on obtient

$$\delta_j'^i = \tau_k^i \sigma_j^l \delta_l^k. \tag{6}$$

Si δ_l^k se définissent par la formule (5), parmi les n^2 termes dans le second membre de (6) sont nuls tous ceux pour lesquels $k \neq l$. Donc, $\delta_j'^i = \tau_k^i \sigma_j^k$. Or $\tau_k^i \sigma_j^k$ est un élément du produit $S^{-1}S$. Il s'ensuit que $\delta_j'^i = \delta_j^i$. On voit que si le tenseur de type (1, 1) possède dans une base la matrice unité, il la conserve dans toute autre base.

EXEMPLE 7. Considérons une fonction $\mathbf{F}(x_1, \ldots, x_q)$ de q vecteurs, linéaire relativement à chacun d'eux si les autres sont fixés. Cette fonction est une généralisation de la forme bilinéaire. Décomposons chacun des vecteurs $x_1, \ldots, x_q$ suivant les vecteurs d'une base quelconque $\boldsymbol{e}$. En vertu de la linéarité par rapport à chacun des vecteurs,

$$\mathbf{F}(x_1, \ldots, x_n) = \mathbf{F}(\xi_1^{i_1} e_{i_1}, \ldots, \xi_q^{i_q} e_{i_q}) = \\ = \xi_1^{i_1} \ldots \xi_q^{i_q} \mathbf{F}(e_{i_1}, \ldots, e_{i_q}) = \alpha_{i_1 \ldots i_q} \xi_1^{i_1} \ldots \xi_q^{i_q},$$

où les coefficients $\alpha_{i_1 \ldots i_q} = \mathbf{F}(e_{i_1}, \ldots, e_{i_q})$ remplissent le même rôle que les coefficients de la forme bilinéaire. Démontrons que dans tout changement de base, ils se transforment comme les composantes du tenseur de type $(0, q)$. A cet effet, considérons la base $e_i' = \sigma_i^k e_k$ et profitons de nouveau de la linéarité de la fonction :

$$\mathbf{F}(e_{i_1}', \ldots, e_{i_q}') = \mathbf{F}(\sigma_{i_1}^{k_1} e_{k_1}, \ldots, \sigma_{i_q}^{k_q} e_{k_q}) = \sigma_{i_1}^{k_1} \ldots \sigma_{i_q}^{k_q} \mathbf{F}(e_{k_1}, \ldots, e_{k_q}).$$

L'égalité obtenue démontre l'assertion énoncée.

D'une façon analogue, on peut construire un exemple de tenseur de type (p, q) quelconque. La fonction $\mathbf{F}$ doit dans ce cas dépendre de q vecteurs et de p covecteurs et être linéaire en chacun d'eux. On peut calculer la valeur de cette fonction sur les vecteurs $x_1, \ldots, x_q$ et sur les covecteurs $l^1, \ldots, l^p$ en décomposant les vecteurs $x_1, \ldots, x_q$ suivant les vecteurs d'une base quelconque $\|e_1, \ldots, e_n\|$ dans $\mathscr{L}_n$ et les covecteurs $l^1, \ldots, l^p$ suivant les vecteurs $\mathbf{p}^1, \ldots, \mathbf{p}^n$ de la base dans l'espace $\mathscr{L}_n^*$, qu'on choisit biorthogonale à la base $\|e_1, \ldots, e_n\|$. Rappelons (comp. p. 224) qu'une base $\|\mathbf{p}^1, \ldots, \mathbf{p}^n\|$ dans l'espace $\mathscr{L}_n^*$ de tous les covecteurs est dite *biorthogonale* à la base $\|e_1, \ldots, e_n\|$ dans $\mathscr{L}_n$ si

$$\mathbf{p}^i(e_k) = \delta_k^i.$$

En vertu de la linéarité de la fonction $\mathbf{F}$, il vient

$$\mathbf{F}(x_1, \ldots, x_q, l^1, \ldots, l^p) = \mathbf{F}(\xi_1^{j_1} e_{j_1}, \ldots, \xi_q^{j_q} e_{j_q}, \lambda_{i_1}^1 \mathbf{p}^{i_1}, \ldots, \lambda_{i_p}^p \mathbf{p}^{i_p}) = \\ = \xi_1^{j_1} \ldots \xi_q^{j_q} \lambda_{i_1}^1 \ldots \lambda_{i_p}^p \mathbf{F}(e_{j_1}, \ldots, e_{j_q}, \mathbf{p}^{i_1}, \ldots, \mathbf{p}^{i_p}) = \\ = \alpha_{j_1 \ldots j_q}^{i_1 \ldots i_p} \xi_1^{j_1} \ldots \xi_q^{j_q} \lambda_{i_1}^1 \ldots \lambda_{i_p}^p.$$

La base $\|\mathbf{p}'^1, \ldots, \mathbf{p}'^n\|$ biorthogonale à la base $\|e_1', \ldots, e_n'\|$ est composée des covecteurs $\mathbf{p}'^j = \tau_i^j \mathbf{p}^i$. En effet,

$$\mathbf{p}'^j(e_l') = \tau_i^j \mathbf{p}^i(\sigma_l^k e_k) = \tau_i^j \sigma_l^k \delta_k^i = \delta_l^j.$$

Il s'ensuit comme plus haut que $\alpha_{j_1 \ldots j_q}^{i_1 \ldots i_p}$ se transforment comme les composantes du tenseur de type (p, q).

Les fonctions étudiées dans cet exemple s'appellent *fonctions multilinéaires*.

Le dernier exemple montre qu'il existe des tenseurs de tout type (p, q). Le tenseur de type (p, q) sera défini si l'on construit une fonction multilinéaire de p covecteurs et q vecteurs. On peut la construire à partir d'une base et d'une matrice de coefficients $\alpha^{i_1 \dots i_p}_{j_1 \dots j_q}$. Après quoi à chaque $(p + q)$-uple de p covecteurs et q vecteurs on peut faire correspondre un nombre

$$\alpha^{i_1 \dots i_p}_{j_1 \dots j_q} \xi^{j_1}_1 \dots \xi^{j_q}_q \lambda^1_{i_1} \dots \lambda^p_{i_p},$$

ce qui définit bien une fonction multilinéaire.

4. Addition et multiplication par un nombre. Pour les tenseurs de même type on définit l'opération d'addition. A savoir, soient $\alpha^{i_1 \dots i_p}_{j_1 \dots j_q}$ et $\beta^{i_1 \dots i_p}_{j_1 \dots j_q}$ les composantes de deux tenseurs **A** et **B** de type (p, q) dans la même base $\boldsymbol{e}$. Associons à cette base une matrice $(p + q)$-dimensionnelle en additionnant les composantes des tenseurs **A** et **B** de mêmes indices :

$$\gamma^{i_1 \dots i_p}_{j_1 \dots j_q} = \alpha^{i_1 \dots i_p}_{j_1 \dots j_q} + \beta^{i_1 \dots i_p}_{j_1 \dots j_q}. \qquad (7)$$

Proposition 3. *Si à chaque base on fait correspondre des nombres* $\gamma^{i_1 \dots i_p}_{j_1 \dots j_q}$ *définis en fonction des composantes des tenseurs* **A** *et* **B** *dans cette base par la formule* (7), *on définit par là même un tenseur de type* (p, q).

Pour le démontrer, il suffit d'établir comment se transforment les nombres $\gamma^{i_1 \dots i_p}_{j_1 \dots j_q}$ lorsqu'on effectue un changement de base. On a

$$\alpha'^{i_1 \dots i_p}_{j_1 \dots j_q} = \tau^{i_1}_{k_1} \dots \tau^{i_p}_{k_p} \sigma^{l_1}_{j_1} \dots \sigma^{l_q}_{j_q} \alpha^{k_1 \dots k_p}_{l_1 \dots l_q}$$

et

$$\beta'^{i_1 \dots i_p}_{j_1 \dots j_q} = \tau^{i_1}_{k_1} \dots \tau^{i_p}_{k_p} \sigma^{l_1}_{j_1} \dots \sigma^{l_q}_{j_q} \beta^{k_1 \dots k_p}_{l_1 \dots l_q}.$$

En additionnant membre à membre ces égalités, on obtient

$$\alpha'^{i_1 \dots i_p}_{j_1 \dots j_q} + \beta'^{i_1 \dots i_p}_{j_1 \dots j_q} = \tau^{i_1}_{k_1} \dots \tau^{i_p}_{k_p} \sigma^{l_1}_{j_1} \dots \sigma^{l_q}_{j_q} (\alpha^{k_1 \dots k_p}_{l_1 \dots l_q} + \beta^{k_1 \dots k_p}_{l_1 \dots l_q}),$$

autrement dit, on obtient la loi tensorielle de transformation pour $\gamma^{i_1 \dots i_p}_{j_1 \dots j_q}$.

Définition. Le tenseur **C** de composantes $\gamma^{i_1 \dots i_p}_{j_1 \dots j_q}$ définies par la formule (7) s'appelle *somme des tenseurs* **A** *et* **B** et se note **A** + **B**.

Ainsi, pour calculer la somme des tenseurs, on additionne leurs composantes correspondantes. Cela signifie que la somme habituelle des vecteurs sera leur somme tensorielle si l'on les additionne comme des tenseurs. Il en est de même de la somme des transformations linéaires, définie dans le chapitre VI. (La définition de la somme y a été donnée pour des applications arbitraires.)

Proposition 4. *Faisons correspondre à chaque base les nombres*

$\lambda\alpha^{i_1\ldots i_p}_{j_1\ldots j_q}$, *où* $\alpha^{i_1\ldots i_p}_{j_1\ldots j_q}$ *sont les composantes du tenseur* **A** *dans cette base. Cette correspondance définit un tenseur.*

On ne donne pas la démonstration de cette proposition, car elle se fait sur le modèle de celle de la proposition 3, où on utilise le fait que les seconds membres des équations (4) sont des polynômes homogènes linéaires par rapport aux anciennes composantes.

DÉFINITION. Le tenseur défini dans la proposition 4 s'appelle *produit du tenseur* **A** *par le nombre* λ et se note λ**A**.

Les propriétés des opérations introduites sont énoncées dans la proposition suivante.

PROPOSITION 5. *L'ensemble de tous les tenseurs de type* (p, q), *muni des opérations d'addition et de multiplication par un nombre est un espace vectoriel de dimension* n^{p+q}.

Laissons au lecteur le soin de vérifier tous les axiomes dans la définition de l'espace vectoriel.

Choisissons une base quelconque dans $\mathscr{L}_n$ et considérons les tenseurs dont l'une des composantes dans la base donnée est égale à l'unité et les autres sont nulles. Il existe exactement n^{p+q} de tels tenseurs car un tenseur quelconque de type (p, q) possède n^{p+q} composantes. Ces tenseurs sont linéairement indépendants et chaque tenseur est leur combinaison linéaire (dont les coefficients sont les composantes du tenseur donné). Donc, la dimension de l'espace des tenseurs de type (p, q) est égale à n^{p+q}.

5. Multiplication des tenseurs. Soient **A** un tenseur de type (p, q) et **B** un tenseur de type (r, s). On peut faire correspondre à une base quelconque $\boldsymbol{e}$ une matrice $(p + q + r + s)$-dimensionnelle composée des produits de chaque composante de **A** par chaque composante de **B**. Ordonnons ces produits en écrivant d'abord les indices se rapportant à **A** puis ceux qui se rapportent à **B**. Il vient

$$\gamma^{i_1\ldots i_p k_1\ldots k_r}_{j_1\ldots j_q l_1\ldots l_s} = \alpha^{i_1\ldots i_p}_{j_1\ldots j_q}\beta^{k_1\ldots k_r}_{l_1\ldots l_s}. \quad (8)$$

PROPOSITION 6. *Si à chaque base on fait correspondre des nombres* $\gamma^{i_1\ldots k}_{j_1\ldots l_s}$ *définis en fonction des composantes des tenseurs* **A** *et* **B** *dans cette base par la formule* (8), *on définit par là même un tenseur de type* $(p + r, q + s)$.

Faisons la démonstration pour le cas des tenseurs de types (1, 1) et (0, 1). Dans le cas général, la démonstration ne diffère que par des notations plus encombrantes. Exprimons les composantes des tenseurs **A** et **B** par rapport à la base $\boldsymbol{e}'$ en fonction de leurs composantes par rapport à la base $\boldsymbol{e}$:

$$\alpha'^{i}_{j} = \tau^{i}_{k}\sigma^{l}_{j}\alpha^{k}_{l}, \quad \beta'_{m} = \sigma^{h}_{m}\beta_{h}.$$

D'où

$$\gamma'^{i}_{jm} = \alpha'^{i}_{j}\beta'_{m} = \tau^{i}_{k}\sigma^{l}_{j}\sigma^{h}_{m}\alpha^{k}_{l}\beta_{h} = \tau^{i}_{k}\sigma^{l}_{j}\sigma^{h}_{m}\gamma^{k}_{lh},$$

autrement dit, les quantités γ^{k}_{lh} se transforment par changement de base comme les composantes du tenseur de type (1, 2).

A la différence des précédentes, cette démonstration utilise l'expression des coefficients affectant les anciennes composantes dans les formules (4) par l'intermédiaire de σ^{i}_{j} et τ^{i}_{j}.

DÉFINITION. Le tenseur défini dans la proposition 6 s'appelle *produit du tenseur* **A** *par le tenseur* **B** et se note $\mathbf{A} \otimes \mathbf{B}$.

EXEMPLE. Considérons deux fonctions linéaires **f** et **h** sur $\mathscr{L}_n$ et faisons correspondre à chaque couple de vecteurs x et y le nombre $\mathbf{f}(x) \cdot \mathbf{h}(y)$. Admettons que dans une certaine base les valeurs de **f** et **h** sont $\mathbf{f}(x) = \varkappa_i \xi^i$ et $\mathbf{h}(y) = \mu_k \eta^k$, où ξ^i et η^k sont les composantes des vecteurs x et y. Dans ce cas

$$\mathbf{b}(x, y) = \mathbf{f}(x) \cdot \mathbf{h}(y) = (\varkappa_i \xi^i)(\mu_k \eta^k) = (\varkappa_i \mu_k)\xi^i \eta^k,$$

car pour calculer le produit de deux polynômes on doit multiplier chaque terme du premier polynôme par chaque terme du second. Ainsi donc, la fonction **b** que nous avons définie est une forme bilinéaire, c'est-à-dire un tenseur de type (0, 2). Ce tenseur est le produit tensoriel des tenseurs **f** et **h**. On peut écrire $\mathbf{b} = \mathbf{f} \otimes \mathbf{h}$, ou en composantes, $\beta_{ik} = \varkappa_i \mu_k$.

La multiplication tensorielle n'est pas commutative. Montrons-le sur l'exemple suivant. Le produit tensoriel de deux vecteurs x et y est le tenseur $x \otimes y$ de type (2, 0). Ses n^2 composantes forment une matrice carrée (le premier indice est celui de la ligne). Si ξ^i et η^k sont les composantes de x et y dans une certaine base, la matrice des composantes de $x \otimes y$ dans cette base est de la forme

$$\begin{Vmatrix} \xi^1\eta^1 & \xi^1\eta^2 & \dots & \xi^1\eta^n \\ \xi^2\eta^1 & \xi^2\eta^2 & \dots & \xi^2\eta^n \\ \dots & \dots & \dots & \dots \\ \xi^n\eta^1 & \xi^n\eta^2 & \dots & \xi^n\eta^n \end{Vmatrix}.$$

En construisant de la même façon la matrice des composantes du tenseur $y \otimes x$, on obtient la matrice transposée. Chaque composante $\xi^i\eta^k$ est le produit de deux nombres, de sorte que $\xi^i\eta^k = \eta^k\xi^i$. Mais l'ordre des composantes des tenseurs $x \otimes y$ et $y \otimes x$ est différent. Il en est de même pour le produit de tenseurs arbitraires.

Laissons au lecteur le soin de démontrer que la multiplication tensorielle est associative et distributive relativement à l'addition. On constate de même aisément que la multiplication d'un tenseur par un nombre, intro-

duite auparavant, coïncide avec la multiplication par le tenseur (0, 0) ayant ce nombre pour composante.

Proposition 7. *Tout tenseur de type (p, q) est une combinaison linéaire de produits dont chacun contient p vecteurs et q covecteurs.*

Il suffit de démontrer que les produits de forme exigée sont les tenseurs qui ont servi dans la proposition 5 à la construction de la base dans l'espace des tenseurs de type (p, q). On le fera pour les tenseurs de type (2, 1) car dans le cas général le raisonnement est le même.

Soit le tenseur **Q** dont la composante θ_1^{23} par rapport à la base $\| e_1, \ldots, e_n \|$ est 1 et les autres composantes sont nulles. Considérons les vecteurs de base e_2, e_3 et le covecteur **f** dont les composantes par rapport à $\| e_1, \ldots, e_n \|$ sont $(1, 0, \ldots, 0)$. Vu que e_2 et e_3 ont pour composantes $(0, 1, 0, \ldots, 0)$ et $(0, 0, 1, 0, \ldots, 0)$, le produit $e_2 \otimes e_3 \otimes \mathbf{f}$ a une seule composante (d'indices 2, 3 et 1) égale à l'unité, et les autres composantes nulles. Donc, $\mathbf{Q} = e_2 \otimes e_3 \otimes \mathbf{f}$. L'assertion se démontre de la même manière pour les autres tenseurs de base dans l'espace des tenseurs de type considéré.

6. Contraction. Soient $\alpha_{j_1 \ldots j_q}^{i_1 \ldots i_p}$ les composantes d'un tenseur de type (p, q) dans la base e, avec $p \geqslant 1$ et $q \geqslant 1$. Choisissons un indice supérieur, par exemple le premier et un indice inférieur, par exemple le dernier. Pour chaque ensemble donné de valeurs des autres indices, considérons la somme de toutes les composantes du tenseur **A** où les valeurs des indices choisis sont les mêmes :

$$\beta_{j_1 \ldots j_{q-1}}^{i_2 \ldots i_p} = \alpha_{j_1 \ldots j_{q-1} 1}^{1 i_2 \ldots i_p} + \alpha_{j_1 \ldots j_{q-1} 2}^{2 i_2 \ldots i_p} + \ldots + \alpha_{j_1 \ldots j_{q-1} n}^{n i_2 \ldots i_p}. \tag{9}$$

Proposition 8. *Faisons correspondre à chaque base une famille de nombres que l'on obtient à partir des composantes $\alpha_{j_1 \ldots j_q}^{i_1 \ldots i_p}$ du tenseur de type (p, q) en appliquant la formule* (9). *Cette correspondance définit un tenseur de type $(p - 1, q - 1)$.*

Pour le démontrer, voyons comment se transforme la famille de nombres ainsi définie, lorsqu'on change de base. On a

$$\beta'^{i_2 \ldots i_p}_{j_1 \ldots j_{q-1}} = \alpha'^{k i_2 \ldots i_p}_{j_1 \ldots j_{q-1} k} = \tau_{m_1}^{k} \tau_{m_2}^{i_2} \ldots \tau_{m_p}^{i_p} \sigma_{j_1}^{l_1} \ldots \sigma_{j_{q-1}}^{l_{q-1}} \sigma_k^{l_q} \alpha_{l_1 \ldots l_q}^{m_1 \ldots m_p}.$$

Or $\tau_{m_1}^{k} \sigma_k^{l_q} = \delta_{m_1}^{l_q}$, de sorte que cette expression est égale à

$$\delta_{m_1}^{l_q} \tau_{m_2}^{i_2} \ldots \tau_{m_p}^{i_p} \sigma_{j_1}^{l_1} \ldots \sigma_{j_{q-1}}^{l_{q-1}} \alpha_{l_1 \ldots l_q}^{m_1 \ldots m_p}.$$

En sommant suivant les indices l_q et m_1, on voit que tous les termes s'annulent, à l'exception de ceux pour lesquels $l_q = m_1$. En posant $l_q = m_1 = k$, on peut écrire

$$\beta'^{i_2 \ldots i_p}_{j_1 \ldots j_{q-1}} = \tau_{m_2}^{i_2} \ldots \tau_{m_p}^{i_p} \sigma_{j_1}^{l_1} \ldots \sigma_{j_{q-1}}^{l_{q-1}} \alpha_{l_1 \ldots l_{q-1} k}^{k m_2 \ldots m_p}.$$

Cela veut dire que les nombres $\beta^{i_2 \dots i_p}_{j_1 \dots j_{q-1}}$ se transforment comme les composantes du tenseur de type $(p - 1, q - 1)$.

DÉFINITION. Le tenseur déduit du tenseur **A** selon la formule (9) s'appelle *tenseur contracté* par rapport au premier indice supérieur et au dernier indice inférieur. On définit de façon analogue le tenseur contracté par rapport à tout indice supérieur et tout indice inférieur.

On appelle *produit contracté de deux tenseurs* un tenseur obtenu par la contraction de leur produit tensoriel, relative à un indice supérieur d'un facteur et à un indice inférieur de l'autre. Par exemple, l'image d'un vecteur x de composantes ξ^i par la transformation linéaire **A** de matrice α^i_j est le produit contracté des tenseurs correspondants : $\eta^i = \alpha^i_j \xi^j = \alpha^i_j \xi^1 + \dots$ $\dots + \alpha^i_n \xi^n$.

La valeur de la fonction linéaire **f** de coefficients $\varkappa_i$ sur le vecteur x de composantes ξ^i est le produit contracté $\zeta = \varkappa_i \xi^i$.

Un exemple significatif de la contraction des indices est donné par le tenseur de type (1, 1). On aboutit au tenseur de type (0, 0), c'est-à-dire à un invariant. Si l'on désigne les composantes d'un tenseur **A** de type (1, 1) par α^i_j, le tenseur contracté obtenu est de la forme $\alpha^i_i = \alpha^1_1 + \dots + \alpha^n_n$. C'est la somme des éléments diagonaux de la matrice A formée des composantes de **A**. On dit que la somme des éléments diagonaux de la matrice A est sa *trace* et on la note tr$A$. On a vu à la page 195 que tr$A$ est un invariant si A est la matrice d'une transformation linéaire. Maintenant on y est arrivé en partant d'un point de vue plus général.

Par contre, pour les tenseurs de type (0, 2), le tenseur contracté n'est pas défini. Il s'ensuit que la trace de la matrice associée à la forme quadratique ne sera pas un invariant, elle peut être modifiée par changement de base. Laissons au lecteur le soin de le vérifier.

7. Transposition. Considérons un ensemble d'éléments de la matrice s-dimensionnelle pour lequel tous les indices, sauf deux, ont des valeurs fixées. Cet ensemble forme une couche bidimensionnelle, autrement dit une matrice carrée. Ainsi, toute la matrice se subdivise en des couches bidimensionnelles qui correspondent aux différentes combinaisons des valeurs de tous les indices, à l'exception de deux choisis.

On appelle *transposition d'une matrice s-dimensionnelle* par rapport à deux indices quelconques une permutation de ses éléments qui transpose chaque couche qu'on obtient en fixant tous les indices à l'exception des deux donnés. Par exemple, la transposition de la matrice $\alpha^{i_1 \dots i_p}_{j_1 \dots j_q}$ par rapport aux deux premiers indices supérieurs donne la matrice $\beta^{i_1 \dots i_p}_{j_1 \dots j_q}$ liée à la première par l'égalité

$$\beta^{i_1 i_2 i_3 \dots i_p}_{j_1 \dots j_q} = \alpha^{i_2 i_1 i_3 \dots i_p}_{j_1 \dots j_q}. \tag{10}$$

D'une façon générale, on entend par transposition suivant un ensemble d'indices le résultat des transpositions successives suivant différents couples d'indices de cet ensemble.

L'opération de transposition est parfois appelée *permutation des indices*.

Considérons à titre d'exemple le cas de $n = 2$. Soit une matrice tridimensionnelle α_{ijk}. Aux valeurs 1 et 2 du premier indice sont associées deux couches. Ecrivons-les l'une à côté de l'autre :

$$\left\| \begin{array}{cc|cc} \alpha_{111} & \alpha_{112} & \alpha_{211} & \alpha_{212} \\ \alpha_{121} & \alpha_{122} & \alpha_{221} & \alpha_{222} \end{array} \right\| .$$

La transposition de cette matrice par rapport à deux derniers indices donne la matrice $\beta_{ijk} = \alpha_{ikj}$, ou sous forme développée

$$\left\| \begin{array}{cc|cc} \beta_{111} & \beta_{112} & \beta_{211} & \beta_{212} \\ \beta_{121} & \beta_{122} & \beta_{221} & \beta_{222} \end{array} \right\| = \left\| \begin{array}{cc|cc} \alpha_{111} & \alpha_{121} & \alpha_{211} & \alpha_{221} \\ \alpha_{112} & \alpha_{122} & \alpha_{212} & \alpha_{222} \end{array} \right\| .$$

Si on considère une transposition plus compliquée $\gamma_{ijk} = \alpha_{kij}$, la matrice γ_{ijk} prend la forme

$$\left\| \begin{array}{cc|cc} \gamma_{111} & \gamma_{112} & \gamma_{211} & \gamma_{212} \\ \gamma_{121} & \gamma_{122} & \gamma_{221} & \gamma_{222} \end{array} \right\| = \left\| \begin{array}{cc|cc} \alpha_{111} & \alpha_{211} & \alpha_{121} & \alpha_{221} \\ \alpha_{112} & \alpha_{212} & \alpha_{122} & \alpha_{222} \end{array} \right\| .$$

PROPOSITION 9. *Supposons qu'à chaque base est associée une matrice $(p + q)$-dimensionnelle $\beta^{i_1 \dots i_p}_{j_1 \dots j_q}$ déduite de la matrice $\alpha^{i_1 \dots i_p}_{j_1 \dots j_q}$ des composantes du tenseur* **A** *par une transposition dans laquelle ne sont permutés que les indices supérieurs ou les indices inférieurs. Cette correspondance définit un tenseur* **B** *de même type que* **A**.

Il suffit de le démontrer pour des transpositions qui ne changent l'ordre que de deux indices, vu que toute transposition est le résultat d'une succession de ces transpositions. La démonstration est de plus identique pour tout couple d'indices supérieurs ou inférieurs. Aussi se limitera-t-on à démontrer la proposition pour une transposition relative à deux premiers indices supérieurs, écrite dans la formule (10). On a

$$\beta'^{i_1 \dots i_p}_{j_1 \dots j_q} = \alpha'^{i_2 i_1 i_3 \dots i_p}_{j_1 \dots j_q} = \tau^{i_2}_{k_1} \tau^{i_1}_{k_2} \tau^{i_3}_{k_3} \dots \tau^{i_p}_{k_p} \sigma^{l_1}_{j_1} \dots \sigma^{l_q}_{j_q} \alpha^{k_1 k_2 k_3 \dots k_p}_{l_1 \dots l_q} .$$

On est en droit de changer la notation des indices de sommation et remplacer k_1 par k_2, et k_2 par k_1, et dans chaque terme permuter les facteurs. On aboutit alors à la formule

$$\beta'^{i_1 i_2 i_3 \dots i_p}_{j_1 \dots j_q} = \tau^{i_1}_{k_1} \tau^{i_2}_{k_2} \tau^{i_3}_{k_3} \dots \tau^{i_p}_{k_p} \sigma^{l_1}_{j_1} \dots \sigma^{l_q}_{j_q} \alpha^{k_2 k_1 k_3 \dots k_p}_{l_1 \dots l_q} =$$
$$= \tau^{i_1}_{k_1} \tau^{i_2}_{k_2} \tau^{i_3}_{k_3} \dots \tau^{i_p}_{k_p} \sigma^{l_1}_{j_1} \dots \sigma^{l_q}_{j_q} \beta^{k_1 k_2 k_3 \dots k_p}_{l_1 \dots l_q} ,$$

qui exprime la loi tensorielle de transformation de $\beta^{i_1 \dots i_p}_{j_1 \dots j_q}$.

DÉFINITION. On dit que le tenseur **B** défini dans la proposition 9 est le *transposé* du tenseur **A**.

Remarquons que la démonstration de la proposition 9 n'aurait pas passée si on avait voulu permuter un indice supérieur avec un indice inférieur. Il en découle par exemple le fait suivant. Etant donné deux matrices A et $S^{-1}AS$ qui définissent une même transformation linéaire dans deux bases différentes, leurs transposées ${}^tA$ et ${}^tS\,{}^tA\,{}^t(S^{-1})$ définissent des transformations linéaires différentes.

Mais si l'on transpose les matrices B et ${}^tSBS$ qui définissent, dans différentes bases, la forme bilinéaire **b**, on aboutit aux matrices ${}^tB$ et ${}^tS\,{}^tBS$ qui, elles aussi, définissent une même (en général, une autre) forme bilinéaire ${}^t\mathbf{b}$. Ces formes bilinéaires sont liées entre elles. On vérifie sans difficulté que ${}^t\mathbf{b}(x, y) = \mathbf{b}(y, x)$.

Les tenseurs qui sont des produits de deux tenseurs donnés pris dans l'ordre différent, se déduisent l'un de l'autre par transposition.

8. Symétrisation et alternation. Considérons le tenseur **A** dont la valence contravariante p est au moins égale à un nombre donné $s \geqslant 2$. Choisissons s indices supérieurs quelconques et calculons le nombre de tenseurs qu'on peut obtenir de **A** par la transposition relative aux indices choisis. En permutant les indices deux à deux on peut les disposer dans un ordre quelconque. Aussi le problème se réduit-il au calcul du nombre de procédés permettant d'ordonner s indices. On peut le faire par $s!$ procédés. Ainsi donc, en transposant le tenseur **A** par rapport à s indices, on peut obtenir $s!$ tenseurs. Additionnons tous ces tenseurs et divisons la somme par $s!$. Le tenseur obtenu est le *symétrisé* du tenseur **A** par rapport aux indices choisis. On désigne ses composantes en mettant entre parenthèses les indices qui affectent les composantes du tenseur **A**. Les indices qui ne participent pas à la symétrisation peuvent être encadrés par des traits verticaux.

On définit de façon analogue la symétrisation par rapport aux indices inférieurs. Donnons quelques exemples :

$$\alpha^{(i|j|k)} = \frac{1}{2}(\alpha^{ijk} + \alpha^{kji}),$$

$$\alpha^i_{(jkl)} = \frac{1}{6}(\alpha^i_{jkl} + \alpha^i_{jlk} + \alpha^i_{klj} + \alpha^i_{kjl} + \alpha^i_{ljk} + \alpha^i_{lkj}).$$

Considérons de nouveau le tenseur **A** de type (p, q), avec $p \geqslant s$, et choisissons s indices supérieurs quelconques. Si l'on numérote ces indices de 1 à s, à chaque tenseur obtenu de **A** par transposition correspondra une permutation $\sigma_1, \ldots, \sigma_s$ des numéros $1, \ldots, s$ (cela veut dire que la transposition de **A** met l'indice de numéro σ_1 à la première place parmi les indices choisis, puis elle met l'indice de numéro σ_2 à la deuxième, etc.). Notons $N(\sigma_1, \ldots$

..., σ_s) le nombre de perturbations de l'ordre dans la permutation σ_1, ..., σ_s, c'est-à-dire le nombre de couples $(\sigma_\alpha, \sigma_\beta)$ tels que $\alpha < \beta$ et $\sigma_\alpha > \sigma_\beta$ (comp. p. 138).

En transposant le tenseur **A** suivant les s indices choisis, on obtient $s!$ tenseurs. Additionnons tous ces tenseurs en multipliant chacun d'eux préalablement par $(-1)^{N(\sigma_1, \ldots, \sigma_s)}$, où σ_1, ..., σ_s est la permutation qui lui est associée. Divisons la somme obtenue par $s!$. On dit que le tenseur ainsi construit est l'*alterné* du tenseur **A** par rapport aux indices choisis. On note ses composantes en mettant entre crochets les indices participant à l'alternation. Exemple :

$$\alpha^{[i|j|k]} = \frac{1}{2}(\alpha^{ijk} - \alpha^{kji}).$$

On définit de façon analogue une alternation relative aux indices inférieurs. Exemple :

$$\alpha^i_{[jkl]} = \frac{1}{6}(\alpha^i_{jkl} + \alpha^i_{ljk} + \alpha^i_{klj} - \alpha^i_{kjl} - \alpha^i_{lkj} - \alpha^i_{jlk}).$$

Donnons encore un exemple. Considérons le déterminant de la matrice d'une transformation linéaire. Dans le § 4 du ch. VI on a constaté que c'est un invariant. Exprimons maintenant cet invariant à l'aide des opérations tensorielles. Soient α^i_j les éléments de la matrice associée à la transformation linéaire **A** dans une base $\boldsymbol{e}$. La n-ième puissance tensorielle $\mathbf{A} \otimes \otimes \mathbf{A} \otimes \ldots \otimes \mathbf{A}$ de la transformation **A** possède alors les composantes $\alpha^{i_1}_{j_1}\alpha^{i_2}_{j_2} \ldots \alpha^{i_n}_{j_n}$. Alternons ce produit par rapport à tous les indices inférieurs, puis contractons par rapport à tous les indices. On obtient l'invariant

$$\Delta = \alpha^{i_1}_{[i_1}\alpha^{i_2}_{i_2} \ldots \alpha^{i_n}_{i_n]}.$$

On a ici n indices de sommation dont chacun prend n valeurs, de sorte que le second membre se subdivise en n^n termes. Chacun de ces termes représente à son tour une somme algébrique de $n!$ termes qui apparaissent par l'alternation. Si dans l'ensemble d'indices déterminant l'une de ces sommes algébriques il y a deux indices égaux, cette somme est nulle. En effet, pour chacun de ces termes pris avec le signe plus, il existe un terme semblable affecté du signe moins. Mais si dans l'ensemble d'indices qui définit la somme algébrique tous les indices sont différents, on peut, en permutant les facteurs dans chacun des termes, la réduire à la forme $\alpha^1_{[1} \ldots \alpha^n_{n]}$. Le nombre total de sommes algébriques étant $n!$, on a

$$\Delta = n!\alpha^1_{[1} \ldots \alpha^n_{n]}.$$

En explicitant l'alternation, on obtient

$$\Delta = n! \frac{1}{n!} \sum_{(k_1 \ldots k_n)} (-1)^{N(k_1 \ldots k_n)} \alpha^1_{k_1} \ldots \alpha^n_{k_n} = \det |\alpha^j_i|.$$

Définition. Un tenseur est dit *symétrique par rapport à un couple d'indices* s'il s'annule par alternation relative à ce couple.

Le tenseur symétrique par rapport à un couple d'indices ne change pas par transposition relative à ce couple. En effet, si par exemple $\alpha_{[ij]} = 0$, on a $\alpha_{ij} = \alpha_{ji}$.

Un tenseur est *symétrique par rapport à un groupe d'indices* s'il l'est pour tout couple d'indices de ce groupe. Il ne change dans ce cas par aucune transposition relative à ce groupe d'indices.

Il n'est pas difficile de se convaincre que la symétrisation du tenseur par rapport à plusieurs indices donne un tenseur symétrique par rapport à ces indices.

Définition. Un tenseur est dit *antisymétrique par rapport à un couple d'indices* s'il devient nul par symétrisation relative à ces indices.

Un tenseur antisymétrique par rapport à un couple d'indices change de signe par transposition relative à ce couple d'indices.

Un tenseur est *antisymétrique par rapport à un groupe d'indices* s'il est antisymétrique par rapport à tout couple d'indices de ce groupe. Dans ce cas, il ne varie par aucune transposition à laquelle correspond une permutation ayant un nombre pair de perturbations d'ordre, et change de signe si la transposition engendre une permutation ayant un nombre impair de perturbations d'ordre. En effet, ces transpositions se réduisent respectivement à un nombre pair et impair de transpositions dont chacune permute un couple d'indices.

On peut démontrer que par alternation d'un tenseur par rapport à plusieurs indices, on peut le rendre antisymétrique relativement à ces indices.

Si un tenseur est antisymétrique par rapport à deux indices, toutes ses composantes avec valeurs confondues de ces indices sont nulles. En effet, soit par exemple $\alpha^{i_1 i_2}_{j_1 j_2 j_3} = -\alpha^{i_2 i_1}_{j_1 j_2 j_3}$. Pour les composantes ayant $i_1 = i_2 = k$, on a alors $\alpha^{kk}_{j_1 j_2 j_3} = -\alpha^{kk}_{j_1 j_2 j_3}$ et, par suite, $\alpha^{kk}_{j_1 j_2 j_3} = 0$.

Démontrons à titre d'exemple la proposition suivante.

Proposition 10. *Chaque tenseur de type* (0, 2) *s'exprime de façon univoque par la somme d'un tenseur symétrique et d'un tenseur antisymétrique.*

En effet, admettons qu'une telle somme existe, c'est-à-dire que $\alpha_{ij} = \beta_{ij} + \gamma_{ij}$, où $\beta_{[ij]} = 0$ et $\gamma_{(ij)} = 0$. Dans ce cas, en alternant et en symétri-

sant les deux membres de l'égalité, on obtient respectivement

$$\alpha_{[ij]} = \gamma_{[ij]} = \gamma_{ij} \quad \text{et} \quad \alpha_{(ij)} = \beta_{(ij)} = \beta_{ij}.$$

D'autre part, on a toujours

$$\alpha_{(ij)} + \alpha_{[ij]} = \frac{1}{2}\alpha_{ij} + \frac{1}{2}\alpha_{ji} + \frac{1}{2}\alpha_{ij} - \frac{1}{2}\alpha_{ji},$$

c'est-à-dire

$$\beta_{ij} + \gamma_{ij} = \alpha_{ij},$$

et la proposition est démontrée.

9. Remarque. Soit une relation entre les tenseurs exprimée à l'aide des opérations tensorielles introduites. Etant donné une base, on peut lui faire correspondre les mêmes relations entre les composantes des tenseurs considérés. Les opérations tensorielles sont invariantes en un sens que les relations entre les composantes sont de la même forme, quelle que soit la base. A savoir, la relation $\mathbf{A} = x \otimes y + z \otimes z$, où x, y et z sont des vecteurs, est équivalente à l'égalité $\alpha^{ij} = \xi^i \eta^j + \zeta^i \zeta^j$ entre les composantes, indifféremment de la nature de la base, vu que dans toutes les bases elle est de la même forme.

En vertu de ce fait, lorsqu'on parle des tenseurs, on a souvent en vue leurs composantes et, inversement, en parlant des composantes, on sous-entend un tenseur. On dit, par exemple, « tenseur α_{ijk} » au lieu de « tenseur dont les composantes dans la base donnée sont α_{ijk} ». On ne crée ainsi aucune ambiguïté et l'on simplifie fortement l'écriture. Dans la suite, on utilisera des abréviations de ce genre.

§ 2. Tenseurs dans l'espace euclidien

1. Tenseur métrique. Tout ce qui a été dit des tenseurs dans un espace vectoriel se rapporte évidemment au cas de l'espace euclidien. Toutefois dans l'espace euclidien, les tenseurs possèdent de nombreuses propriétés qu'ils n'avaient pas dans l'espace vectoriel.

Le théorème 2 du § 3, ch. VIII, peut être formulé en termes des tenseurs de la façon suivante : un espace euclidien de dimension n peut être défini comme un espace vectoriel de dimension n dans lequel est donné un tenseur symétrique $\mathbf{G}$ de composantes g_{ij}, avec $g_{ij}\xi^i\xi^j > 0$ pour tout vecteur ξ^i non nul. Les composantes du tenseur $\mathbf{G}$ dans la base e sont les éléments de la matrice de Gram de cette base.

Définition. Associons à chaque base de l'espace euclidien la matrice de Gram de cette base. Le tenseur de type (0, 2) défini par cette correspondance s'appelle *tenseur métrique* de l'espace euclidien.

PROPOSITION 1. *Faisons correspondre à chaque base de l'espace euclidien la matrice inverse de la matrice de Gram de cette base. Cette correspondance définit un tenseur de type* (2, 0).

DÉMONSTRATION. Soient Γ et Γ' les matrices de Gram des bases e et $e' = eS$. Voyons comment sont liées les matrices Γ^{-1} et $(\Gamma')^{-1}$. On s'appuiera sur la formule (15) du § 1, ch. VII : $\Gamma' = {}^t S \Gamma S$. Cherchons les matrices inverses de Γ' et ${}^t S \Gamma S$ en nous servant de la formule (10) du § 6, ch. V :

$$(\Gamma')^{-1} = S^{-1}({}^t S \Gamma)^{-1} = S^{-1}\Gamma^{-1}({}^t S)^{-1}.$$

Comme on l'a vu au § 6 du ch. V, pour toute matrice de déterminant non nul on a $({}^t S)^{-1} = {}^t(S^{-1})$. Donc,

$$(\Gamma')^{-1} = (S^{-1})(\Gamma^{-1})\,{}^t(S^{-1}). \tag{1}$$

Ecrivons la formule (1) en désignant par g^{ij}, g'^{ij} et τ^i_j les éléments des matrices Γ^{-1}, $(\Gamma')^{-1}$ et S^{-1}. Il vient,

$$g'^{ij} = \tau^i_k \tau^j_l g^{kl},$$

autrement dit, on a la loi de transformation des composantes du tenseur de type (2, 0).

DÉFINITION. Le tenseur construit dans la proposition 1 s'appelle *tenseur métrique contravariant*.

Les égalités $\Gamma\Gamma^{-1} = E$ et $\Gamma^{-1}\Gamma = E$ peuvent être écrites au moyen des composantes g_{ij} et g^{ij} ainsi :

$$g_{ij}g^{jk} = \delta^k_i, \quad g_{ji}g^{kj} = \delta^k_i.$$

En outre, vu que ${}^t(\Gamma^{-1}) = ({}^t\Gamma)^{-1} = \Gamma^{-1}$, le tenseur métrique contravariant est symétrique :

$$g^{ij} = g^{ji}.$$

2. Elévation et abaissement des indices. Le tenseur métrique permet d'introduire dans l'espace euclidien deux opérations sur les tenseurs : l'élévation et l'abaissement des indices.

Abaissement de l'indice. C'est une opération qui permet, à partir d'un tenseur de type (p, q), $p \geqslant 1$, d'obtenir un tenseur de type $(p - 1, q + 1)$. Cette opération combine la multiplication du tenseur donné et du tenseur métrique avec la contraction par rapport à l'indice qu'on veut abaisser. Dans ce cas, l'ordre des indices se conserve au sens suivant. On n'observe plus la convention suivant laquelle chaque indice inférieur suit chaque indice supérieur. Pour noter l'ordre des indices on met un point au-dessus de chaque indice inférieur et au-dessous de chaque indice supérieur. Par exemple, en abaissant le premier indice du tenseur $\alpha^{ij\cdot}_{\cdot\cdot k}$, on obtient le tenseur $g_{il}\alpha^{lj\cdot}_{\cdot\cdot k} = \alpha^{\cdot j \cdot}_{l \cdot k}$.

Elévation de l'indice. Cette opération permet, à partir d'un tenseur de type (p, q), d'obtenir un tenseur de type $(p + 1, q - 1)$. Elle combine la multiplication du tenseur donné et du tenseur métrique contravariant avec la contraction relative à l'indice qu'on veut élever. Il faut bien entendu que le tenseur de départ ait au moins un indice inférieur, c'est-à-dire que $q \geqslant 1$. Par exemple, l'élévation du premier indice de $\alpha_{l.k}^{\cdot j\cdot}$ donne $\alpha^{ij}_{\cdot\cdot k} = g^{il}\alpha_{l.k}^{\cdot j\cdot}$, et l'élévation du troisième, $\alpha_{l\cdot\cdot}^{\cdot jk} = g^{kl}\alpha_{i.l}^{\cdot j\cdot}$.

On a déjà rencontré l'opération d'abaissement de l'indice.

Dans le § 2 du ch. VII, on a introduit la notion de transformation linéaire adjointe **A*** de la transformation linéaire donnée **A** de l'espace euclidien. Dans une base arbitraire, les matrices A et A^* des transformations **A** et **A*** sont liées par l'égalité (3) du § 2, ch. VII, qu'on peut mettre sous la forme $\Gamma A^* = {}^t(\Gamma A)$. En désignant par α^i_j et α^{*i}_j les éléments des matrices A et A^*, écrivons cette égalité en notations tensorielles :

$$g_{ik}\alpha^k_j = g_{jk}\alpha^{*k}_i.$$

Comme on le voit aisément, elle veut dire que l'abaissement des indices des tenseurs **A** et **A*** donne deux tenseurs dont l'un se déduit de l'autre par transposition. Pour une transformation symétrique **A**, le tenseur $g_{ik}\alpha^k_j$ est symétrique.

Dans le § 3 du ch. VIII, on a introduit la notion de transformation linéaire **A** associée à la forme bilinéaire **b**. Dans une base arbitraire, la matrice A de la transformation associée s'exprime au moyen de la matrice B de la forme bilinéaire par l'égalité $A = \Gamma^{-1}B$. En notations tensorielles on peut l'écrire ainsi :

$$\alpha^i_j = g^{ik}\beta_{kj}.$$

On voit que le tenseur **A** se déduit du tenseur **b** par élévation du premier indice.

Dans le § 1 du ch. VIII a été introduit un vecteur associé à la fonction linéaire sur l'espace euclidien. La matrice-ligne $\boldsymbol{x}$ des coefficients de cette fonction s'exprimait au moyen de la matrice-colonne $\boldsymbol{\alpha}$ des coordonnées du vecteur par la formule $\boldsymbol{x} = {}^t\boldsymbol{\alpha}\Gamma$. En passant de la représentation matricielle à l'écriture au moyen des composantes, on obtient $x_i = \alpha^k g_{ki}$. Il en découle que $\alpha^j = g^{ij}x_i$ (ch. VIII).

3. Tenseurs euclidiens. En étudiant les espaces euclidiens, on peut souvent se limiter aux seules bases orthonormées. Dans ce cas, toutes les formules contenant des produits scalaires se simplifient fortement car le tenseur métrique a une matrice unité

$$g_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{2}$$

Dans tout changement de base orthonormée, la matrice de passage est orthogonale, c'est-à-dire satisfait à la relation $S^{-1} = {}^tS$. Cela veut dire que les éléments de S et de son inverse sont liés par les égalités

$$\sigma_k^i = \tau_i^k. \tag{3}$$

Si l'on introduit la relation (3) dans la loi de transformation des composantes du tenseur (4), § 1, on obtient

$$\alpha'^{i_1 \dots i_p}_{j_1 \dots j_q} = \sum_{\substack{k_1, \dots, k_p \\ l_1, \dots, l_q}} \sigma_{i_1}^{k_1} \dots \sigma_{i_p}^{k_p} \sigma_{j_1}^{l_1} \dots \sigma_{j_q}^{l_q} \alpha^{k_1 \dots k_p}_{l_1 \dots l_q}. \tag{4}$$

Les indices de sommation $k_1, \dots, k_p$ se disposent ici en haut, c'est pourquoi on écrit le symbole de sommation. La formule (4) montre qu'en se limitant aux bases orthonormées, on élimine la différence entre les indices supérieurs et inférieurs : aux indices supérieurs correspondent les mêmes facteurs dans la loi de transformation (4) qu'aux indices inférieurs.

Signalons de plus qu'en vertu de (2) tous les tenseurs dont les indices diffèrent de hauteur ont mêmes composantes dans une base orthonormée donnée. En effet, on a, par exemple, $\alpha_{lij} = g_{ik} \alpha^{k\cdot}_{l\cdot j} = \alpha^{i\cdot}_{l\cdot j}$, vu que de n termes de la somme suivant k n'est différent de zéro que celui pour lequel $i = k$. Or on a pour ce dernier $g_{ii} = 1$.

Il ressort de ce qu'on vient d'exposer qu'en se limitant à l'étude des bases orthonormées, on peut identifier tous les tenseurs déduits l'un de l'autre par élévation ou abaissement de l'indice. Plus précisément, tous les tenseurs qui possèdent les mêmes composantes par rapport à une base orthonormée sont réunis en une seule classe qu'on étudie comme un nouvel objet. Cet objet est le tenseur euclidien.

Définition. On dit que dans l'espace euclidien $\mathcal{E}_n$ est défini un *tenseur euclidien de valence s* si à chaque base orthonormée est associée une matrice s-dimensionnelle d'ordre n. Ceci étant, quelles que soient les bases orthonormées e et e', les éléments $\alpha_{i_1 \dots i_s}$ et $\alpha'_{i_1 \dots i_s}$ des matrices respectives sont liés par la relation

$$\alpha'_{i_1 \dots i_s} = \sigma_{i_1}^{k_1} \dots \sigma_{i_s}^{k_s} \alpha_{k_1 \dots k_s}. \tag{5}$$

Tous les indices d'un tenseur euclidien jouent le même rôle, si bien qu'on les écrit tous en bas.

Il découle de la formule (3) que tout tenseur de type (p, q) définit un tenseur euclidien de valence $(p + q)$. Ceci étant, tous les tenseurs différant l'un de l'autre par la hauteur d'un indice définissent un même tenseur euclidien.

Etant donné un tenseur euclidien, ses composantes ne sont définies dans aucune base non orthonormée. Mais pour tout tenseur euclidien de valence s, ces composantes peuvent être déterminées de manière que s'obtienne un tenseur de type (p, q), où $p + q = s$. Il faut pour cela les rechercher en utilisant la loi de transformation des composantes d'un tenseur de type adéquat. Ainsi, chaque tenseur euclidien est engendré par tout tenseur d'une certaine classe. Il en ressort facilement que tous les tenseurs de cette classe ne diffèrent que par la hauteur de l'indice.

Les grandeurs numériques ne variant pas par changement de base orthonormée ont été appelées à la p. 248 *invariants orthogonaux* (ou *euclidiens*). On voit maintenant que ce sont des tenseurs euclidiens de valence 0.

On peut définir pour les tenseurs euclidiens toutes les opérations introduites pour les tenseurs au § 1. Les définitions ainsi que les énoncés et démonstrations des propositions correspondantes reproduiront presque mot à mot ce qui a été déjà dit, aussi s'abstiendra-t-on de les donner. Signalons seulement que pour les tenseurs euclidiens, la contraction s'avère possible relativement à tout couple d'indices, et la transposition peut s'effectuer par rapport à tout ensemble d'indices. Par exemple, si, en se limitant aux bases orthonormées, on identifie la

forme quadratique à la transformation qui lui est associée, le nouvel objet obtenu (le tenseur euclidien de valence 2) sera invariant par contraction (comme la transformation linéaire) et vérifiera dans toute base orthonormée l'égalité $\alpha_{ij} = \alpha_{ji}$ (comme la forme quadratique).

§ 3. Multivecteurs. Invariants relatifs

1. *p*-vecteurs. On débutera dans ce paragraphe par l'étude d'une classe spéciale de tenseurs.

DÉFINITION. Un tenseur de type $(p, 0)$, antisymétrique par rapport à tous les indices est appelé *p-vecteur*, ou *multivecteur* si p n'est pas précisé.

Un tenseur de type $(0, q)$, antisymétrique par rapport à tous les indices est appelé *q-covecteur* ou *q-forme*.

La théorie des q-covecteurs est analogue à la théorie des multivecteurs. On aura affaire essentiellement aux p-vecteurs.

Remarquons tout d'abord que pour $p > n$ il n'existe qu'un p-vecteur nul. En effet, parmi $p > n$ indices prenant les valeurs $1, \ldots, n$ deux au moins doivent avoir des valeurs égales. Donc, dans chaque composante il y a deux indices égaux. On sait que la composante doit changer de signe dans la transposition du tenseur permutant ces indices. Or, comme les indices sont égaux, elle ne varie pas et, par suite, vaut zéro. Le raisonnement peut être répété pour chaque composante.

Pour $p = n$, ne peuvent être différentes de zéro que les composantes dont les valeurs des indices constituent une permutation des nombres $1, \ldots$ $\ldots, n$ (autrement on aurait de nouveau deux indices égaux). Toutes ces composantes s'expriment au moyen d'une d'entre elles par la formule

$$\alpha^{i_1 \ldots i_n} = (-1)^{N(i_1, \ldots, i_n)} \alpha^{1 \ldots n}, \tag{1}$$

où $N(i_1, \ldots, i_n)$ est le nombre de perturbations de l'ordre dans la permutation $i_1, \ldots, i_n$ (comp. p. 138).

En effet, on va montrer que les valeurs des indices $i_1, \ldots, i_n$ peuvent être rangées dans l'ordre naturel par $N(i_1, \ldots, i_n)$ transpositions permutant deux indices, et que chacune de ces transpositions modifie le signe de la composante. On peut ranger les valeurs des indices dans l'ordre naturel de la façon suivante : l'un des nombres $i_1, \ldots, i_n$ étant égal à n, déplaçons-le à la dernière place en le permutant successivement avec tous les nombres qui se trouvent à sa droite ; avec chacun de ces nombres n a engendré une perturbation de l'ordre. On effectuera donc autant de permutations que n a engendré de perturbations sans changer l'ordre des autres nombres. Prenons maintenant le nombre $n - 1$ et déplaçons-le de la même manière à l'avant-dernière place. On fera ainsi autant de permutations de nombres pris deux à deux qu'il y a de perturbations engendrées par $n - 1$. On ne changera toujours pas l'ordre des autres nombres. Agissons de même avec

les nombres $n-2, \ldots, 2$. Après quoi tous les nombres, dont 1 qui automatiquement prend la première place, se disposeront dans l'ordre naturel. Pour y arriver, on a fait $N(i_1, \ldots, i_n)$ permutations de nombres pris deux à deux.

Un p-vecteur $\alpha^{i_1 \ldots i_p}$ est dit *simple* s'il est un produit alterné de vecteurs, c'est-à-dire

$$\alpha^{i_1 \ldots i_p} = \xi_1^{[i_1} \ldots \xi_p^{i_p]}, \tag{2}$$

où, par exemple, $\xi_p^{i_p}$ sont les composantes du vecteur x_p.

PROPOSITION 1. *Tout n-vecteur est simple.*

DÉMONSTRATION. Considérons un n-vecteur simple non nul $\beta^{i_1 \ldots i_n} = \xi_1^{[i_1} \ldots \xi_n^{i_n]}$. Désignons par λ le rapport de la composante $\alpha^{1 \ldots n}$ du n-vecteur donné à $\beta^{1 \ldots n}$. Il ressort alors de la formule (1) qu'on a aussi pour toutes les composantes $\alpha^{i_1 \ldots i_n} = \lambda \beta^{i_1 \ldots i_n}$. En posant $\eta^i = \lambda \xi_1^i$, on a $\alpha^{i_1 \ldots i_n} = \eta^{[i_1} \xi_2^{i_2} \ldots \xi_n^{i_n]}$, ce qu'il fallait démontrer.

Soient donnés p vecteurs $x_1, \ldots, x_p$ de composantes $\xi_1^{i_1}, \ldots, \xi_p^{i_p}$.

Composons la matrice dont les colonnes sont les colonnes de coordonnées des vecteurs $x_1, \ldots, x_p$:

$$\begin{Vmatrix} \xi_1^1 & \ldots & \xi_p^1 \\ \xi_1^2 & \ldots & \xi_p^2 \\ \ldots & \ldots & \ldots \\ \xi_1^n & \ldots & \xi_p^n \end{Vmatrix}. \tag{3}$$

Choisissons p lignes de cette matrice de numéros $i_1 < i_2 < \ldots < i_p$ et considérons le mineur situé sur ces lignes :

$$M^{i_1 \ldots i_p} = \sum_{\sigma_1, \ldots, \sigma_p} (-1)^{N(i_{\sigma_1}, \ldots, i_{\sigma_p})} \xi_1^{i_{\sigma_1}} \ldots \xi_p^{i_{\sigma_p}},$$

où $\sigma_1, \ldots, \sigma_p$ sont tous différents et prennent les valeurs $1, \ldots, p$. En comparant cette expression à (2), on obtient les composantes du p-vecteur simple dont les valeurs de tous les indices sont différentes deux à deux et se disposent dans l'ordre naturel :

$$\alpha^{i_1 \ldots i_p} = \xi_1^{[i_1} \ldots \xi_p^{i_p]} = \frac{1}{p!} M^{i_1 \ldots i_p}. \tag{4}$$

Les autres composantes soit s'annulent (s'il y a des indices égaux), soit diffèrent des composantes déjà trouvées par le facteur $(-1)^N$, où N est le nombre de perturbations dans l'ordre des valeurs d'indices.

En particulier, on a pour le n-vecteur

$$\alpha^{1\ldots n} = \frac{1}{n!}\begin{vmatrix} \xi_1^1 & \cdots & \xi_n^1 \\ \ldots & \ldots & \ldots \\ \xi_1^n & \cdots & \xi_n^n \end{vmatrix}. \tag{5}$$

Il ressort de la formule (4) que le p-vecteur simple est nul si et seulement si les vecteurs qui le composent sont linéairement dépendants, c'est-à-dire si le rang de la matrice (3) est strictement inférieur à p.

2. Invariants relatifs. Tout n-vecteur se caractérise complètement dans chaque base par un nombre, sa composante $\alpha^{1\ldots n}$. Essayons d'écrire la loi de transformation de ce nombre dans un changement de base, sans faire intervenir les autres composantes du n-vecteur. On obtient

$$\alpha'^{1\ldots n} = \tau^1_{i_1} \ldots \tau^n_{i_n}\alpha^{i_1\ldots i_n} = \sum_{(i_1,\ldots,i_n)} (-1)^{N(i_1,\ldots,i_n)}\tau^1_{i_1} \ldots \tau^n_{i_n}\alpha^{1\ldots n},$$

ou

$$\alpha'^{1\ldots n} = (\det S^{-1})\alpha^{1\ldots n} = (\det S)^{-1}\alpha^{1\ldots n}. \tag{6}$$

On voit que la correspondance attachant à chaque base une composante indépendante du n-vecteur donné détermine un objet géométrique défini par une composante et régi par la loi de transformation (6). En effet, à chaque base est associé un nombre unique de telle sorte que le nombre correspondant à une base s'exprime par le nombre correspondant à une autre base et par la matrice de passage. Cet objet n'est pas un tenseur, car un tenseur à une composante est nécessairement de type (0, 0) et, partant, est un invariant.

Un autre exemple d'objet géométrique de ce genre a été mentionné au début du § 1 : le déterminant de la matrice des composantes d'un tenseur de type (0, 2) se transforme d'après la loi

$$\delta' = (\det S)^2\delta. \tag{7}$$

Définition. On dit qu'un *invariant relatif de poids r* est défini dans l'espace vectoriel si à chaque base on peut associer un nombre de façon que les nombres α et α' correspondant aux bases $\boldsymbol{e}$ et $\boldsymbol{e}' = \boldsymbol{e}S$ soient liés par l'égalité

$$\alpha' = (\det S)^r\alpha. \tag{8}$$

Un *invariant*, ou *invariant absolu* pour souligner la différence avec un invariant relatif, est un invariant relatif de poids 0.

Par analogie à la formule (6), on peut montrer que la composante indépendante de tout n-covecteur est un invariant relatif de poids 1.

Mentionnons les propriétés suivantes des opérations sur les invariants relatifs.

PROPOSITION 2. 1) *Soient donnés deux invariants relatifs a et b d'un même poids r. En associant à chaque base la somme des composantes de a et b dans cette base, on obtient un invariant relatif de poids r.*

2) *Soient donnés les invariants relatifs a et b de poids respectifs r_1 et r_2. En associant à chaque base le produit des composantes de a et b dans cette base, on obtient un invariant relatif de poids $r_1 + r_2$.*

3) *Soit un invariant relatif a de poids r. En associant à chaque base la p-ième puissance de la composante de a dans cette base, on obtient un invariant relatif de poids pr.*

Les trois assertions se démontrent de la même façon. Démontrons à titre d'exemple la deuxième. En multipliant les égalités (8) écrites pour a et b, on obtient la relation

$$a'b' = (\det S)^{r_1+r_2}ab$$

entre les produits des composantes de a et b dans deux bases quelconques $\boldsymbol{e}$ et $\boldsymbol{e}'$, ou la loi de transformation recherchée.

On peut maintenant s'assurer qu'il existe des invariants relatifs de tout poids r. Pour construire un invariant relatif de poids r, il suffit d'élever à la puissance r un invariant relatif quelconque de poids 1.

3. Volume d'un parallélépipède *n*-dimensionnel. Considérons dans un espace tridimensionnel ordinaire un parallélépipède engendré par les vecteurs $\mathbf{x}_1$, $\mathbf{x}_2$ et $\mathbf{x}_3$ issus de l'origine des coordonnées. Il peut être défini comme un ensemble des vecteurs de la forme $\mathbf{a} = \lambda_1\mathbf{x}_1 + \lambda_2\mathbf{x}_2 + \lambda_3\mathbf{x}_3$, avec $0 \leqslant \lambda_1 \leqslant 1$, $0 \leqslant \lambda_2 \leqslant 1$ et $0 \leqslant \lambda_3 \leqslant 1$. La formule (10) du § 3, ch. I, fournit le volume du parallélépipède, qui est égal à la valeur absolue du déterminant

$$\begin{vmatrix} \xi_1^1 & \xi_2^1 & \xi_3^1 \\ \xi_1^2 & \xi_2^2 & \xi_3^2 \\ \xi_1^3 & \xi_2^3 & \xi_3^3 \end{vmatrix}$$

construit avec les composantes des vecteurs $\mathbf{x}_1$, $\mathbf{x}_2$ et $\mathbf{x}_3$ dans une base orthonormée. Ce déterminant est égal au produit de 3! par la composante du 3-vecteur engendré par ces vecteurs, dont les indices sont 1, 2 et 3.

DÉFINITION. Etant donné un espace vectoriel de dimension n, on appelle *parallélépipède p-dimensionnel* *) l'ensemble de tous les vecteurs de la forme $a = \lambda_1 x_1 + \ldots + \lambda_p x_p$, où $x_1, \ldots, x_p$ sont des vecteurs linéairement indépendants donnés, et $0 \leqslant \lambda_\sigma \leqslant 1$ pour tous les $\sigma = 1, \ldots, p$.

*) On utilise de même le mot « parallélotope ».

DÉFINITION. Soit le parallélépipède n-dimensionnel P construit sur les vecteurs $x_1, \ldots, x_n$ dans un espace euclidien de dimension n. Si $\xi_1^{i_1}, \ldots, \xi_n^{i_n}$ sont les composantes de ces vecteurs dans une base orthonormée, on appellera *volume* du parallélépipède n dimensionnel le nombre

$$V(P) = n!\,|\xi_1^{[1} \ldots \xi_n^{n]}|, \tag{9}$$

c'est-à-dire le produit de $n!$ par le module de la composante du n-vecteur engendré par les vecteurs $x_1, \ldots, x_n$, dont les indices sont $1, \ldots, n$.

On ne précise pas dans la définition de quelle base orthonormée il s'agit. Aussi doit-on démontrer que $V(P)$ ne dépend pas du choix de la base orthonormée. Selon la formule (6), le passage d'une base orthonormée $\boldsymbol{e}$ à la base $\boldsymbol{e}' = \boldsymbol{e}S$ entraîne $V'(P) = |\det S^{-1}|\, V(P)$. Si $\boldsymbol{e}'$ est orthonormée, la matrice S^{-1} est orthogonale et $|\det S^{-1}| = 1$, ce qui démontre l'assertion.

On voit que $V(P)$ est un invariant orthogonal, c'est-à-dire un tenseur euclidien de valence 0 (voir p. 248).

Proposons-nous d'obtenir l'expression du volume du parallélépipède n-dimensionnel dans une base arbitraire. Remarquons pour cela que selon la formule (7), le déterminant de la matrice du tenseur métrique $\det \Gamma$ est un invariant relatif de poids 2. En outre, on a toujours $\det \Gamma > 0$ (proposition 3, § 1, ch. VII). Aussi peut-on envisager la quantité $\sqrt{\det \Gamma}$ qu'on doit évidemment multiplier par $|\det S|$ quand on change de base. Le produit $n!\sqrt{\det \Gamma}\,|\xi_1^{[1} \ldots \xi_n^{n]}|$ est un invariant absolu. Or dans un base orthonormée, $\det \Gamma = 1$, de sorte que, dans cette base on a

$$V(P) = n!\sqrt{\det \Gamma}\,|\xi_1^{[1} \ldots \xi_n^{n]}|. \tag{10}$$

Le premier membre de l'égalité est un invariant et, par suite, l'égalité (10) est vérifiée dans toute base.

CHAPITRE XI

APPLICATIONS LINÉAIRES

§ 1. Application adjointe

On sait que dans un espace euclidien, toute transformation linéaire **A** admet une transformation adjointe $\mathbf{A}^*$. La même notation avec astérisque a été utilisée pour l'espace vectoriel $\mathcal{L}^*$ dual de l'espace vectoriel donné $\mathcal{L}$, qui est l'ensemble des fonctions linéaires définies sur $\mathcal{L}$. Cette ressemblance de notations traduit un lien profond entre les notions mentionnées. Dans ce paragraphe, on décrira ce lien de façon plus détaillée. L'objet de ce paragraphe est de définir et d'étudier une application adjointe de l'application linéaire donnée d'un espace vectoriel dans l'autre.

1. Orthogonalité. Introduisons quelques notations et notions pour les fonctions linéaires sur un espace vectoriel $\mathcal{L}_n$. Elles vont rappeller le lien existant entre les fonctions linéaires et les vecteurs d'un espace euclidien et simplifier les énoncés de certaines propositions.

Considérons une fonction linéaire f définie sur l'espace vectoriel $\mathcal{L}_n$. La valeur de la fonction f sur le vecteur x sera noté $\langle f, x \rangle$ et non pas $f(x)$. Cette notation paraît plus symétrique pour f et x.

On sait que le vecteur $x \in \mathcal{L}_n$ peut être interprété comme une fonction linéaire sur $\mathcal{L}_n^*$, faisant correspondre à tout f de $\mathcal{L}_n^*$ un nombre $\langle f, x \rangle$ *). Puisque la valeur de f sur x et la valeur de x sur f est un même nombre, on est en droit d'écrire $\langle f, x \rangle = \langle x, f \rangle$.

Le vecteur $x \in \mathcal{L}_n$ sera dit *orthogonal* au vecteur f de $\mathcal{L}_n^*$ si $\langle f, x \rangle = 0$. Notons que la notion de bases biorthogonales suppose une orthogonalité considérée dans ce sens-là. Il est évident que la relation d'orthogonalité est symétrique : f est orthogonal à x si et seulement si x est orthogonal à f.

Définition. Soit $\mathcal{M}_k$ un sous-espace vectoriel dans $\mathcal{L}_n$. L'ensemble de tous les vecteurs de $\mathcal{L}_n^*$ orthogonaux à tout vecteur de $\mathcal{M}_k$ est appelé l'*orthogonal* de $\mathcal{M}_k$ et est noté $\mathcal{M}_k^{\perp}$.

Soulignons que l'orthogonal d'un sous-espace de $\mathcal{L}_n$ se trouve dans un autre espace, l'espace $\mathcal{L}_n^*$.

---

*) Pour le lecteur initié à l'algèbre tensoriel, il est peut-être utile de noter que $\langle f, x \rangle$ est le produit contracté du vecteur x et du covecteur f.

Proposition 1. *L'orthogonal du sous-espace k-dimensionnel* $\mathscr{M}_k$ *est un sous-espace vectoriel avec* dim $\mathscr{M}_k^{\perp} = n - k$.

La démonstration ne diffère presque pas de celle de la proposition analogue énoncée pour l'espace euclidien. Remarquons d'abord que $f \in \mathscr{M}_k^{\perp}$ si et seulement si f est orthogonal à tous les vecteurs d'une base quelconque dans $\mathscr{M}_k$. En effet, si $\|a_1, \ldots, a_k\|$ est une base dans $\mathscr{M}_k$, pour tout x de $\mathscr{M}_k$ on a $x = \xi^1 a_1 + \ldots + \xi^k a_k$, de sorte que $\langle f, x \rangle = \xi^1 \langle f, a_1 \rangle + \ldots$ $\ldots + \xi^k \langle f, a_k \rangle$ en vertu de la linéarité de la fonction f. Par conséquent, si $\langle f, a_1 \rangle = 0, \ldots, \langle f, a_k \rangle = 0$, on a $\langle f, x \rangle = 0$. Inversement, si $\langle f, x \rangle = 0$ pour tous les $x \in \mathscr{M}_k$, il en est de même en particulier pour les vecteurs de base.

L'espace $\mathscr{L}_n$ étant rapporté à une base, on est en mesure d'exprimer en coordonnées les égalités $<f, a_i> = 0$ pour tous les $i = 1, \ldots, k$. On obtient un système d'équations linéaires

$$\begin{aligned} x_1 \alpha_1^1 + \ldots + x_n \alpha_1^n &= 0, \\ \ldots\ldots\ldots&\ldots\ldots\ldots \\ x_1 \alpha_k^1 + \ldots + x_n \alpha_k^n &= 0 \end{aligned}$$

par rapport aux coefficients de la fonction f dans la base choisie. $\alpha_i^1, \ldots, \alpha_i^n$ désignent les composantes du vecteur a_i. Vu que les vecteurs $a_1, \ldots, a_k$ sont linéairement indépendants, le rang de la matrice du système vaut k. On sait que l'ensemble des solutions du système a une structure d'un espace vectoriel de dimension $n - k$, ce qui équivaut à l'assertion qu'il fallait démontrer.

Considérons pour un sous-espace $\mathscr{M}_k$ l'espace $(\mathscr{M}_k^{\perp})^{\perp}$ qui est l'orthogonal de son orthogonal. La relation d'orthogonalité étant inversible, il s'ensuit que $\mathscr{M}_k \subseteq (\mathscr{M}_k^{\perp})^{\perp}$. Or la dimension de $(\mathscr{M}_k^{\perp})^{\perp}$ est égale à $n - (n - k) = k$. Cela signifie que $(\mathscr{M}_k^{\perp})^{\perp}$ se confond avec $\mathscr{M}_k$.

2. Définition de l'application adjointe. Considérons deux espaces vectoriels $\mathscr{L}_n$ et $\bar{\mathscr{L}}_m$ de dimensions n et m, en général, différentes. $\mathscr{L}_n^*$ et $\bar{\mathscr{L}}_m^*$ désignent comme d'habitude les espaces duals correspondants. Soit donnée une application linéaire $\mathbf{A} : \mathscr{L}_n \to \bar{\mathscr{L}}_m$.

Si $f \in \bar{\mathscr{L}}_m^*$ est une fonction linéaire sur $\bar{\mathscr{L}}_m$, on est en mesure de définir le produit de l'application $\mathbf{A}$ par cette fonction, c'est-à-dire d'agir d'abord sur le vecteur $x \in \mathscr{L}_n$ par l'application $\mathbf{A}$ et, ensuite, de prendre la valeur de la fonction f sur le vecteur $\mathbf{A}(x) \in \bar{\mathscr{L}}_m$. On définit ainsi une fonction numérique g sur l'espace $\mathscr{L}_n$, qui peut être notée $f \circ \mathbf{A}$. Ainsi donc,

$$\langle g, x \rangle = (f \circ \mathbf{A})(x) = \langle f, \mathbf{A}(x) \rangle . \tag{1}$$

La fonction f peut être interprétée comme une application linéaire de $\bar{\mathscr{L}}_m$ dans l'ensemble de nombres et, par suite, g est linéaire en tant que produit

de deux applications linéaires. D'ailleurs, la linéarité de g peut être vérifiée directement.

Associons maintenant à chaque fonction f sur $\mathscr{L}_m$ une fonction $f \circ \mathbf{A}$ sur $\mathscr{L}_n$ et notons $\mathbf{A}^*$ une application ainsi obtenue de $\mathscr{L}_m^*$ dans $\mathscr{L}_n^*$. Ainsi,

$$\mathbf{A}^*(f) = f \circ \mathbf{A}. \tag{2}$$

Considérons la valeur de chacun des membres de l'égalité (2) sur le vecteur x. Il vient,

$$\langle \mathbf{A}^*(f), x \rangle = \langle f, \mathbf{A}(x) \rangle . \tag{3}$$

L'égalité (3) ressemble beaucoup à la définition de la transformation adjointe dans l'espace euclidien. On montrera plus loin que ce n'est pas seulement une ressemblance. Remarquons que la ressemblance des expressions à été obtenue grâce aux notations convenablement choisies. Si la valeur de la fonction était notée $f(x)$, le second membre de l'égalité (3) aurait dû être écrit ainsi : $f(\mathbf{A}(x))$, et le premier membre, sous une forme absolument incommode : $\mathbf{A}^*(f)(x)$.

Démontrons que l'application $\mathbf{A}^*$ est linéaire. A cet effet, rappelons les définitions de la somme de deux fonctions linéaires et du produit d'une fonction linéaire par un nombre. Dans les nouvelles notations, ces définitions prennent la forme

$$\langle f_1 + f_2, x \rangle = \langle f_1, x \rangle + \langle f_2, x \rangle ,$$
$$\langle \alpha f, x \rangle = \alpha \langle f, x \rangle .$$

En se servant de ces égalités et de l'égalité (3), on peut écrire

$$\langle \mathbf{A}^*(f_1 + f_2), x \rangle = \langle f_1 + f_2, \mathbf{A}(x) \rangle = \langle f_1, \mathbf{A}(x) \rangle + \langle f_2, \mathbf{A}(x) \rangle =$$
$$= \langle \mathbf{A}^*(f_1), x \rangle + \langle \mathbf{A}^*(f_2), x \rangle = \langle \mathbf{A}^*(f_1) + \mathbf{A}^*(f_2), x \rangle$$

et

$$\langle \mathbf{A}^*(\alpha f), x \rangle = \langle \alpha f, \mathbf{A}(x) \rangle = \alpha \langle f, \mathbf{A}(x) \rangle = \alpha \langle \mathbf{A}^*(f), x \rangle ,$$

ce qu'il fallait démontrer. On peut maintenant introduire la définition suivante.

DÉFINITION. L'application linéaire $\mathbf{A}^* : \mathscr{L}_m^* \to \mathscr{L}_n^*$ définie par la formule (2) est dite ***adjointe*** de l'application linéaire $\mathbf{A} : \mathscr{L}_n \to \mathscr{L}_m$.

On aurait pu interpréter la formule (3) comme une définition en exigeant que l'égalité soit vérifiée pour tout $x \in \mathscr{L}_n$.

Il ressort de ce qui vient d'être dit que toute application linéaire admet une application adjointe et une seule.

Considérons l'application adjointe $\mathbf{A}^{**}$ de l'application $\mathbf{A}^*$. Selon la définition, $\mathbf{A}^{**}$ applique $\mathscr{L}_n$ dans $\mathscr{L}_m$. Ceci étant, si l'on interprète $x \in \mathscr{L}_n$

comme une fonction linéaire sur $\mathscr{L}_n^*$, on obtient pour tout $f \in \mathscr{L}_m^*$ l'égalité suivante

$$\langle \mathbf{A}^{**}(x), f \rangle = \langle x, \mathbf{A}^*(f) \rangle .$$

Comme il a été indiqué plus haut, on peut mettre cette égalité sous une autre forme :

$$\langle f, \mathbf{A}^{**}(x) \rangle = \langle \mathbf{A}^*(f), x \rangle ,$$

d'où on obtient, en appliquant la définition à $\mathbf{A}^*$, que

$$\langle f, \mathbf{A}^{**}(x) \rangle = \langle f, \mathbf{A}(x) \rangle .$$

Puisque cette égalité est vérifiée pour tout $f \in \mathscr{L}_m^*$, les vecteurs $\mathbf{A}^{**}(x)$ et $\mathbf{A}(x)$, considérés comme des fonctions sur $\mathscr{L}_m^*$, doivent coïncider. Leur coïncidence pour tous les $x \in \mathscr{L}_n$ signifie que

$$\mathbf{A}^{**} = \mathbf{A}.$$

3. Expression analytique. Rapportons les espaces $\mathscr{L}_n$ et $\mathscr{L}_m$ aux bases respectives $\boldsymbol{e} = \| e_1, \ldots, e_n \|$ et $\boldsymbol{l} = \| l_1, \ldots, l_m \|$. L'image du vecteur $x \in \mathscr{L}_n$ par l'application $\mathbf{A}$ se définit alors dans la base $\boldsymbol{l}$ par la colonne de coordonnées

$$\boldsymbol{\eta} = A\boldsymbol{\xi},$$

où $\boldsymbol{\xi}$ est la colonne de coordonnées de x par rapport à la base $\boldsymbol{e}$, et A la matrice de l'application $\mathbf{A}$ par rapport aux bases choisies. Considérons une fonction linéaire f sur $\mathscr{L}_m$ et désignons la matrice-ligne de ses coefficients dans la base $\boldsymbol{l}$ par $\boldsymbol{\varkappa} = \| \varkappa_1, \ldots, \varkappa_m \|$. La valeur de la fonction $f \circ \mathbf{A}$ sur le vecteur x est alors

$$\langle f, \mathbf{A}(x) \rangle = \boldsymbol{\varkappa}\boldsymbol{\eta} = \boldsymbol{\varkappa} A \boldsymbol{\xi},$$

de sorte que la fonction $\mathbf{A}^*(f) = f \circ \mathbf{A}$ se définit dans la base $\boldsymbol{e}$ par la matrice-ligne de coordonnées

$$\boldsymbol{\mu} = \boldsymbol{\varkappa} A. \tag{4}$$

Il reste à se rappeler que les coefficients de la fonction linéaire dans une base sont ses coordonnées par rapport à la base de l'espace dual, biorthogonale à la base donnée. Si l'on écrit les coordonnées du vecteur de l'espace dual sous forme de matrice-colonne, la formule (4) devient ${}^t\boldsymbol{\mu} = {}^tA\,{}^t\boldsymbol{\varkappa}$. Il s'ensuit la

PROPOSITION 2. *Si A est la matrice de l'application $\mathbf{A} : \mathscr{L}_n \to \mathscr{L}_m$ par rapport à un couple de bases données, la matrice transpcsée ${}^tA$ est celle de l'application adjointe $\mathbf{A}^* : \mathscr{L}_m^* \to \mathscr{L}_n^*$ par rapport à un couple de bases biorthogonales.*

Si l'on s'en tient au point de vue matriciel, on pourrait dire que la matrice A de type (m, n), traitée comme matrice d'une application, peut être utilisée de deux façons :

1) On peut la multiplier à droite par une matrice-colonne à n éléments et obtenir une matrice-colonne à m éléments. On définit ainsi une application $\mathbf{A}$ de l'espace des matrices-colonnes à n éléments dans l'espace des matrices-colonnes à m éléments.

2) On peut la multiplier à gauche par une matrice-ligne à m éléments et obtenir une matrice-ligne à n éléments. On définit par là-même une application $\mathbf{A}^*$ de l'espace des matrices-lignes à m éléments dans l'espace des matrices-lignes à n éléments.

4. Propriétés des applications adjointes. Le passage à l'application adjointe est lié de la façon suivante aux opérations algébriques sur les applications.

PROPOSITION 3. *Si les applications* $\mathbf{BA}$, $\mathbf{A}^{-1}$ *ou* $\mathbf{A} + \mathbf{B}$ *sont définies, on a respectivement*

$$(\mathbf{BA})^* = \mathbf{A}^* \mathbf{B}^*, \tag{5}$$

$$(\mathbf{A}^{-1})^* = (\mathbf{A}^*)^{-1}, \tag{6}$$

$$(\mathbf{A} + \mathbf{B})^* = \mathbf{A}^* + \mathbf{B}^*, \tag{7}$$

$$(\alpha \mathbf{A})^* = \alpha \mathbf{A}^*. \tag{8}$$

Pour démontrer la formule (5), considérons trois espaces vectoriels $\mathscr{L}$, $\mathscr{L}'$ et $\mathscr{L}''$ et les applications $\mathbf{A} : \mathscr{L} \to \mathscr{L}'$ et $\mathbf{B} : \mathscr{L}' \to \mathscr{L}''$. Si $\mathscr{L}$, $\mathscr{L}'$ et $\mathscr{L}''$ sont rapportés respectivement aux bases $\boldsymbol{e}$, $\boldsymbol{l}$ et $\boldsymbol{h}$, la matrice de l'application $\mathbf{B} \circ \mathbf{A}$ par rapport au couple de bases $\boldsymbol{e}$ et $\boldsymbol{h}$ est égale à BA, où A est la matrice de $\mathbf{A}$ par rapport aux bases $\boldsymbol{e}$, $\boldsymbol{l}$ et B, la matrice de $\mathbf{B}$ par rapport aux bases $\boldsymbol{l}$, $\boldsymbol{h}$. On sait que ${}^t(BA) = {}^tA\,{}^tB$ pour toutes matrices A et B dont le produit BA est défini. Il s'ensuit que la matrice de l'application $(\mathbf{B} \circ \mathbf{A})^*$ par rapport aux bases biorthogonales $\boldsymbol{h}$ et $\boldsymbol{e}$ est ${}^tA\,{}^tB$ et, par suite, $(\mathbf{B} \circ \mathbf{A})^*$ est égale au produit $\mathbf{A}^* \mathbf{B}^*$. La formule (5) est donc démontrée.

Si l'application $\mathbf{A}$ est inversible et A est sa matrice par rapport aux bases $\boldsymbol{e}$ et $\boldsymbol{l}$, l'application inverse $\mathbf{A}^{-1}$ possède la matrice A^{-1} par rapport aux bases $\boldsymbol{l}$ et $\boldsymbol{e}$. Vu que $({}^tA)^{-1} = {}^t(A^{-1})$, l'application adjointe $\mathbf{A}^*$ est également inversible et sa matrice est ${}^t(A^{-1})$.

On ne s'arrêtera pas à la démonstration des formules (7) et (8). Elles découlent facilement de l'expression analytique des applications.

Rappelons qu'on appelle *noyau* de l'application $\mathbf{A}$ l'ensemble des vecteurs x tels que $\mathbf{A}(x) = o$. On désigne le noyau de $\mathbf{A}$ par le symbole Ker $\mathbf{A}$. En utilisant la proposition 2 du § 3, ch. VI, on démontre facilement que Ker $\mathbf{A}$ est un sous-espace vectoriel dans $\mathscr{L}_n$.

PROPOSITION 4. *L'orthogonal de l'ensemble des valeurs de l'application* **A** *se confond avec le noyau de son application adjointe :*

$$(\mathbf{A}(\mathscr{L}_n))^{\perp} = \operatorname{Ker} \mathbf{A}^*. \tag{9}$$

Le noyau Ker $\mathbf{A}^*$ de l'application $\mathbf{A}^*$ est par définition composé de tous les vecteurs $f \in \mathscr{L}_m^*$ tels que $\mathbf{A}^*(f) = o$. Cela signifie que la relation $f \in \operatorname{Ker} \mathbf{A}^*$ est équivalente à l'égalité $\langle \mathbf{A}^*(f), x \rangle = 0$ pour tout $x \in \mathscr{L}_n$, qui, à son tour, est équivalente à $\langle f, \mathbf{A}(x) \rangle = 0$. Or l'orthogonal du sous-espace $\mathbf{A}(\mathscr{L}_n)$ est justement l'ensemble de tous les $f \in \mathscr{L}_m^*$ tels que $\langle f, \mathbf{A}(x) \rangle = 0$ pour tout $x \in \mathscr{L}_n$.

COROLLAIRE. *En appliquant la proposition* 4 *à l'application* $\mathbf{A}^*$ *on obtient*

$$\mathbf{A}^*(\mathscr{L}_m^*) = (\operatorname{Ker} \mathbf{A})^{\perp}.$$

Exprimée en coordonnées, la proposition 4 se réduit au théorème de Fredholm sur la compatibilité des systèmes d'équations linéaires. Rappelons son énoncé.

THÉORÈME DE FREDHOLM. *Un système de m équations linéaires à n inconnues* $A\xi = b$ *est compatible si et seulement si toute solution du système homogène transposé* ${}^tA\eta = \mathbf{0}$ *satisfait à la condition* $\eta b = 0$.

Considérons deux espaces $\mathscr{L}_n$ et $\mathscr{L}_m$ dont chacun est rapporté à une base. Chaque matrice-colonne dans l'énoncé du théorème sera assimilée à une colonne de coordonnées d'un certain vecteur, et la matrice du système A à la matrice de l'application linéaire **A**. La compatibilité du système $A\xi = b$ signifie que $b \in \mathbf{A}(\mathscr{L}_n)$. L'ensemble des solutions du système homogène transposée est Ker $\mathbf{A}^*$. Ceci noté, on constate facilement que l'assertion du théorème de Fredholm s'écrit à l'aide de la formule (9).

PROPOSITION 5. Rg $\mathbf{A}^*$ = Rg **A**.

Cela résulte directement de la proposition 2. Cependant, on recommande au lecteur de déduire cette proposition à titre d'exercice, en comparant les dimensions des deux membres de l'égalité (9).

5. Transformation adjointe. On a défini plus haut une application adjointe pour toute application linéaire d'un espace vectoriel dans l'autre. Considérons maintenant l'application adjointe d'une transformation $\mathbf{A}: \mathscr{L}_n \to \mathscr{L}_n$, c'est-à-dire d'une application pour laquelle les espaces $\mathscr{L}_n$ et $\mathscr{L}_m$ coïncident. Dans ce cas, les espaces $\mathscr{L}_n^*$ et $\mathscr{L}_m^*$ se confondent de même et l'application adjointe est une transformation de l'espace $\mathscr{L}_n^*$. Ainsi, deux transformations adjointes **A** et $\mathbf{A}^*$ opèrent dans deux espaces duals $\mathscr{L}_n$ et $\mathscr{L}_n^*$.

Rappelons que la matrice de l'application $\mathbf{A}: \mathscr{L}_n \to \mathscr{L}_m$ est définie

quand sont choisies les bases : e dans $\mathscr{L}_n$ et l dans $\mathscr{L}_m$. Dans le cas d'une transformation, les bases e et l doivent par définition coïncider : la matrice de la transformation **A** par rapport à la base e est composée des coordonnées des vecteurs $\mathbf{A}(e_1), \ldots, \mathbf{A}(e_n)$ par rapport à e. Cette définition prise en compte, on obtient la

PROPOSITION 6. *Si A est la matrice de la transformation* **A** *par rapport à la base e, ${}^tA$ est celle de la transformation adjointe* $\mathbf{A}^*$ *par rapport à la base p biorthogonale à e.*

Pour les transformations linéaires, à la différence d'applications linéaires quelconques, on peut définir les sous-espaces invariants, les valeurs propres et les vecteurs propres. Etudions comment ils sont liés pour une transformation **A** et son adjoint $\mathbf{A}^*$.

Le polynôme caractéristique ne varie pas par transposition de la matrice : $\det(A - \lambda E) = \det({}^tA - \lambda E)$. De plus, $\operatorname{Rg}(A - \lambda E) = \operatorname{Rg}({}^tA - \lambda E)$, de sorte que la proposition 6 entraîne la

PROPOSITION 7. *Les valeurs propres des transformations* **A** *et* $\mathbf{A}^*$ *se confondent, et les valeurs propres égales sont de même multiplicité. Si à une valeur propre λ de la transformation* **A** *sont associés k vecteurs linéairement indépendants, il en est de même pour la transformation* $\mathbf{A}^*$.

PROPOSITION 8. *Si un sous-espace $\mathscr{M} \subseteq \mathscr{L}_n$ est invariant pàr la transformation* **A**, *son orthogonal $\mathscr{M}^\perp \subseteq \mathscr{L}_n^*$ est invariant par la transformation* $\mathbf{A}^*$.

DÉMONSTRATION. Soit $x \in \mathscr{M}$. Alors $\mathbf{A}(x) \in \mathscr{M}$ et tout vecteur $f \in \mathscr{M}^\perp$ vérifie la relation $\langle f, \mathbf{A}(x) \rangle = 0$. Or cela signifie que $\langle \mathbf{A}^*(f), x \rangle = 0$ et, par suite, $\mathbf{A}^*(f) \in \mathscr{M}^\perp$.

PROPOSITION 9. *Soient x et f les vecteurs propres des transformations respectives* **A** *et* $\mathbf{A}^*$ *associés aux valeurs propres λ et μ. Si $\lambda \neq \mu$, les vecteurs x et f sont orthogonaux.*

En effet, soit

$$\mathbf{A}(x) = \lambda x, \quad x, \mathbf{A}(x) \in \mathscr{L}_n,$$

et

$$\mathbf{A}^*(f) = \mu f, \quad f, \mathbf{A}^*(f) \in \mathscr{L}_n^*.$$

Considérons la valeur de la fonction linéaire f sur le vecteur $\mathbf{A}(x)$. On a

$$\langle f, \mathbf{A}(x) \rangle = \langle \mathbf{A}^*(f), x \rangle = \mu \langle f, x \rangle .$$

Or $\langle f, \mathbf{A}(x) \rangle = \lambda \langle f, x \rangle$. Donc,

$$(\lambda - \mu) \langle f, x \rangle = 0,$$

ce qui démontre la proposition

PROPOSITION 10. *Si la base e est composée des vecteurs propres de la transformation* **A**, *la base biorthogonale à e est composée des vecteurs propres de la transformation* **A***.

Pour le démontrer, il suffit de se rappeler que la matrice de la transformation **A** est diagonale par rapport à la base *e*. Par suite, est diagonale sa matrice transposée. Or c'est la matrice de la transformation **A*** par rapport à la base biorthogonale à *e*.

6. Cas d'espaces euclidiens. Soient $\mathscr{E}_n$ et $\tilde{\mathscr{E}}_m$ des espaces euclidiens de dimensions respectives n et m. Considérons une application linéaire **A** : $\mathscr{E}_n \to \tilde{\mathscr{E}}_m$ et son application adjointe **A***. La propriété essentielle pour nous de l'espace euclidien est qu'il peut être identifié avec son espace dual (voir point 4, § 1, ch. VIII).

Ainsi, pour toute application **A** : $\mathscr{E}_n \to \tilde{\mathscr{E}}_m$ son adjointe **A*** applique $\tilde{\mathscr{E}}_m$ dans $\mathscr{E}_n$ et peut être définie par l'égalité

$$(\mathbf{A}^*(y), x) = (y, \mathbf{A}(x)), \tag{10}$$

qui doit être vérifiée pour tous vecteurs $x \in \mathscr{E}_n$ et $y \in \tilde{\mathscr{E}}_m$. Les parenthèses intervenant dans le premier et le second membre de l'égalité (10) renferment le produit scalaire respectivement dans $\mathscr{E}_n$ et dans $\tilde{\mathscr{E}}_m$. Dans la suite, on s'abstiendra de telles indications en admettant que l'appartenance des facteurs à un espace permet de comprendre que l'opération de multiplication scalaire est envisagée dans cet espace (comme c'est admis pour le signe +).

Remarquons que si **A** est une transformation dans l'espace $\mathscr{E}_n$, son adjointe **A*** est aussi une transformation dans $\mathscr{E}_n$, et la définition (10) coïncide avec la définition connue de la transformation adjointe dans un espace euclidien.

Exprimons la formule (10) sous forme analytique. Rapportons $\mathscr{E}_n$ et $\tilde{\mathscr{E}}_m$ à deux bases quelconques et désignons par $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ les colonnes de coordonnées des vecteurs $x \in \mathscr{E}_n$ et $y \in \tilde{\mathscr{E}}_m$. Soient A et A^* les matrices des applications **A** et **A***, et Γ et $\tilde{\Gamma}$ les matrices de Gram des bases choisies. L'égalité (10) prend alors la forme

$$^t(A^*\boldsymbol{\eta})\Gamma\boldsymbol{\xi} = {}^t\boldsymbol{\eta}\tilde{\Gamma}A\boldsymbol{\xi}.$$

Les matrices $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ étant arbitraires, il s'ensuit que

$$^t(A^*)\,\Gamma = \tilde{\Gamma}A,$$

ou

$$A^* = (\Gamma^{-1})(^tA)\tilde{\Gamma}. \tag{11}$$

Cette expression est beaucoup plus compliquée que la relation entre les matrices, mentionnée dans la proposition 6. On l'explique par le fait qu'en

indentifiant $\mathscr{E}_n$ à $\mathscr{E}_n^*$ et $\tilde{\mathscr{E}}_m$ à $\tilde{\mathscr{E}}_m^*$, on choisit dans les espaces identifiés une même base et non pas deux bases biorthogonales. Si la base est orthonormée, elle coïncide avec sa base biorthogonale. Pour les bases orthonormées, (11) devient $A^* = {}^tA$, ce qui est conforme à la proposition 6.

Comparé au cas général étudié dans les points précédents le cas d'espaces euclidiens contient un fait nouveau : en identifiant $\mathscr{E}_n$ à $\mathscr{E}_n^*$ et $\tilde{\mathscr{E}}_m$ à $\tilde{\mathscr{E}}_m^*$, on définit par là même les produits $\mathbf{A}^*\mathbf{A}$ et $\mathbf{A}\mathbf{A}^*$. Le premier d'entre eux est une transformation de l'espace $\mathscr{E}_n$ et le second, de l'espace $\tilde{\mathscr{E}}_m$. Les propriétés de ces transformations sont semblables. En effet, la transformation $\mathbf{A}\mathbf{A}^*$ peut être exprimée sous la forme $(\mathbf{A}^*)^*(\mathbf{A}^*)$ et, par suite, la seconde des transformations se confond avec la première, prise pour l'application $\mathbf{A}^*$. Etudions les propiétés de $\mathbf{A}^*\mathbf{A}$.

Le rôle principal dans cette étude revient à la relation évidente suivante :

$$(\mathbf{A}^*\mathbf{A}(x), x) = (\mathbf{A}(x), \mathbf{A}(x)) \geqslant 0, \tag{12}$$

qui s'annule si et seulement si $\mathbf{A}(x) = o$.

Proposition 11. *La transformation* $\mathbf{A}^*\mathbf{A}$ *est symétrique. Ses valeurs propres sont positives.*

En effet, selon la proposition 3, $(\mathbf{A}^*\mathbf{A})^* = \mathbf{A}^*\mathbf{A}^{**} = \mathbf{A}^*\mathbf{A}$, ce qui démontre la première assertion. Posons $\mathbf{A}^*\mathbf{A}(x) = \lambda x$. Alors $(\mathbf{A}^*\mathbf{A}(x), x) = \lambda(x, x) \geqslant 0$ en vertu de l'égalité (12). D'où $\lambda \geqslant 0$.

Proposition 12. *Le noyau de la transformation* $\mathbf{A}^*\mathbf{A}$ *se confond avec le noyau de l'application* $\mathbf{A}$. *L'ensemble des valeurs de* $\mathbf{A}^*\mathbf{A}$ *coïncide avec celui de* $\mathbf{A}^*$.

Démonstration. $\mathbf{A}(x) = o$ entraîne $\mathbf{A}^*\mathbf{A}(x) = o$ et, par suite, $\operatorname{Ker} \mathbf{A} \subseteq \operatorname{Ker} \mathbf{A}^*\mathbf{A}$. Par ailleurs, $\mathbf{A}^*\mathbf{A}(x) = o$ entraîne $(\mathbf{A}^*\mathbf{A}(x), x) = 0$ et, en vertu de (12), $\mathbf{A}(x) = o$. Donc, $\operatorname{Ker} \mathbf{A}^*\mathbf{A} \subseteq \operatorname{Ker} \mathbf{A}$, ce qui démontre l'assertion

$$\operatorname{Ker} \mathbf{A} = \operatorname{Ker} \mathbf{A}^*\mathbf{A}. \tag{13}$$

Pour démontrer la seconde assertion, comparons les rangs de $\mathbf{A}$ et de $\mathbf{A}^*\mathbf{A}$, ce qui présente un intérêt en soi. On sait que la somme du rang d'une application et de la dimension de son noyau vaut la dimension de l'espace appliqué. L'espace $\mathscr{E}_n$ est le même pour les applications $\mathbf{A}$ et $\mathbf{A}^*\mathbf{A}$, et leurs noyaux se confondent. Par suite,

$$\operatorname{Rg} \mathbf{A}^*\mathbf{A} = \operatorname{Rg} \mathbf{A}. \tag{14}$$

Ensuite, puisque $\mathbf{A}(\mathscr{E}_n) \subseteq \tilde{\mathscr{E}}_m$, il est évident que $\mathbf{A}^*\mathbf{A}(\mathscr{E}_n) \subseteq \mathbf{A}^*(\tilde{\mathscr{E}}_m)$. De par la définition du rang d'une application on a $\operatorname{Rg} \mathbf{A} = \operatorname{Rg} \mathbf{A}^* = \dim \mathbf{A}^*(\tilde{\mathscr{E}}_m)$ et $\operatorname{Rg} \mathbf{A}^*\mathbf{A} = \dim \mathbf{A}^*\mathbf{A}(\mathscr{E}_n)$. Les rangs étant égaux, les

sous-espaces se confondent en vertu de l'inclusion mentionnée plus haut et de l'égalité des dimensions :

$$\mathbf{A}^* \mathbf{A}(\mathscr{E}_n) = \mathbf{A}^* (\tilde{\mathscr{E}}_m).$$

La proposition est démontrée.

Dans un espace euclidien, à toute transformation symétrique $\mathbf{B}$ on associe une forme quadratique $(\mathbf{B}(x), x)$. De la proposition 11 découle le corollaire suivant.

COROLLAIRE. *La forme quadratique* $(\mathbf{A}^* \mathbf{A}(x), x)$ *est semi-définie positive. Elle est définie positive si et seulement si* $\operatorname{Ker} \mathbf{A} = o$.

On sait que $\operatorname{Rg} \mathbf{A} = \operatorname{Rg} \mathbf{A}^*$. D'où, en appliquant la formule (14) à l'application $\mathbf{A}^*$, on obtient

$$\operatorname{Rg} \mathbf{A}^* \mathbf{A} = \operatorname{Rg} \mathbf{A}\mathbf{A}^*. \qquad (15)$$

Etudions les relations entre les vecteurs propres et les valeurs propres des transformations $\mathbf{A}^* \mathbf{A}$ et $\mathbf{A}\mathbf{A}^*$. Elles s'expriment par la proposition suivante.

PROPOSITION 13. *Si x est un vecteur propre de la transformation* $\mathbf{A}^* \mathbf{A}$ *associé à la valeur propre* $\lambda \neq 0$, $\mathbf{A}(x)$ *est un vecteur propre de la transformation* $\mathbf{A}\mathbf{A}^*$, *associé à la même valeur propre. De plus, aux vecteurs propres linéairement indépendants* $x_1, \ldots, x_s$ *de la transformation* $\mathbf{A}^* \mathbf{A}$ *correspondent les vecteurs propres linéairement indépendants* $\mathbf{A}(x_1), \ldots, \mathbf{A}(x_s)$ *de la transformation* $\mathbf{A}\mathbf{A}^*$.

DÉMONSTRATION. Soit $\mathbf{A}^* \mathbf{A}(x) = \lambda x$. Considérons les images par l'application $\mathbf{A}$ des deux membres de cette équation, soit : $(\mathbf{A}\mathbf{A}^*)\mathbf{A}(x) = \lambda \mathbf{A}(x)$. Si $\lambda \neq 0$, on a $\mathbf{A}^* \mathbf{A}(x) \neq o$ et, par suite, $\mathbf{A}(x) \neq o$. Dans ce cas, $\mathbf{A}(x)$ est un vecteur propre de $\mathbf{A}\mathbf{A}^*$, associé à la valeur λ.

Considérons ensuite les vecteurs propres linéairement indépendants $x_1, \ldots, x_s$ de la transformation $\mathbf{A}^* \mathbf{A}$, associés aux valeurs propres non nulles $\lambda_1, \ldots, \lambda_s$ dont quelques-unes peuvent être égales. Supposons que les vecteurs $\mathbf{A}(x_1), \ldots, \mathbf{A}(x_s)$ sont linéairement dépendants et que $\alpha_1 \mathbf{A}(x_1) + \ldots + \alpha_s \mathbf{A}(x_s)$ est leur combinaison linéaire non triviale égale à zéro. Considérons l'image de cette dernière par l'application $\mathbf{A}^*$. Il vient

$$\alpha_1 \mathbf{A}^* \mathbf{A}(x_1) + \ldots + \alpha_s \mathbf{A}^* \mathbf{A}(x_s) = \alpha_1 \lambda_1 x_1 + \ldots + \alpha_s \lambda_s x_s = o,$$

ce qui contredit l'indépendance linéaire des vecteurs $x_1, \ldots, x_s$ car $\lambda_1, \ldots, \lambda_s$ sont non nuls. La proposition est démontrée.

Vu que pour les transformations symétriques le nombre maximal de vecteurs propres linéairement indépendants associés à la valeur propre λ est égal à sa multiplicité, on obtient le

COROLLAIRE. *Les valeurs propres non nulles des transformations* **A*** **A** *et* **AA*** *coïncident, et de plus les valeurs propres égales sont de même multiplicité.*

REMARQUE. La présence de la valeur propre nulle et sa multiplicité se déterminent pour les transformations **A*** **A** et **AA*** d'après les dimensions de Ker **A** et Ker **A*** respectivement.

7. Bases singulières d'une application. Considérons une application **A** : $\mathscr{E}_n \to \tilde{\mathscr{E}}_m$, où $\mathscr{E}_n$ et $\tilde{\mathscr{E}}_m$ sont deux espaces euclidiens.

DÉFINITION. On appelle *première base singulière* de l'application **A** une base orthonormée dans $\mathscr{E}_n$, composée des vecteurs propres de la transformation **A*** **A**, ordonnés de manière que les valeurs propres correspondantes forment une suite décroissante : $\lambda_1 \geqslant \ldots \geqslant \lambda_n$.

Ainsi donc, si $r = \mathrm{Rg}\, \mathbf{A}$, on a $\lambda_i > 0$ pour $i \leqslant r$, et $\lambda_j = 0$ pour $j > r$. Soit $\|e_1, \ldots, e_n\|$ la première base singulière de **A**. Alors

$$(\mathbf{A}(e_i), \mathbf{A}(e_j)) = (\mathbf{A}^* \mathbf{A}(e_i), e_j) = \lambda_i(e_i, e_j).$$

Il s'ensuit que les vecteurs $\mathbf{A}(e_i)$ sont deux à deux orthogonaux et

$$|\mathbf{A}(e_i)| = \sqrt{\lambda_i}, \tag{16}$$

ce qui signifie que $\mathbf{A}(e_i) \neq o$ pour $i \leqslant r$, et $\mathbf{A}(e_i) = o$ pour $i > r$ *).

DÉFINITION. Les nombres $\alpha_i = \sqrt{\lambda_i}$, où λ_i sont les valeurs propres de la transformation **A*** **A**, s'appellent *nombres singuliers* de l'application **A**, ainsi que nombres singuliers de la matrice de cette application par rapport à toute base.

Selon (16), les vecteurs $\alpha_i^{-1}\mathbf{A}(e_i)$ constituent pour $i \leqslant r$ un système orthonormé dans $\tilde{\mathscr{E}}_m$. Complétons-le jusqu'à une base orthonormée f dans $\tilde{\mathscr{E}}_m$ et introduisons la définition qui suit.

DÉFINITION. On appelle *seconde base singulière* de l'application **A** une base orthonormée f dans $\tilde{\mathscr{E}}_m$ dont les r premiers vecteurs sont de la forme $\alpha_i^{-1}\mathbf{A}(e_i)$, $i = 1, \ldots, r$, où $\|e_1, \ldots, e_n\|$ est la première base singulière et $r = \mathrm{Rg}\, \mathbf{A}$.

Il découle de la définition que les bases singulières ne sont pas définies de façon univoque.

THÉORÈME 1. *Rapportée à un couple de bases singulières de l'application* **A**, *la matrice de cette application est de la forme*

$$A = \begin{Vmatrix} D_r & O \\ O & O \end{Vmatrix}, \tag{17}$$

*) Voir proposition 12.

où D_r est la matrice diagonale carrée d'ordre r avec les nombres α_i sur la diagonale, les autres éléments de A étant nuls.

DÉMONSTRATION. Par définition, les colonnes de A sont les matrices-colonnes des coordonnées des vecteurs $\mathbf{A}(e_i)$ par rapport à la base f. Il ressort de (16) que les $n - r$ dernières colonnes de la matrice A sont nulles. La définition de f entraîne que $\mathbf{A}(e_i) = \alpha_i f_i$ pour $i \leqslant r$. Ce qui achève la démonstration.

Rappelons que dans tout changement de bases dans les espaces $\mathscr{E}_n$ et $\mathscr{E}_m$ la matrice de l'application se transforme suivant la formule $A' = P^{-1}AQ$, où P et Q sont des matrices de passage. Si les bases de départ sont orthonormées, les matrices de passage aux bases singulières seront orthogonales. Aussi peut-on exprimer le théorème 1 en termes de matrices de la façon suivante :

THÉORÈME 1M. *Pour toute matrice A de type* (m, n) *il existe deux matrices orthogonales U et V telles que la matrice UAV soit de la forme* (17).

Un énoncé équivalent est souvent fort utile :

Toute matrice A de type (m, n) peut être décomposée en produit

$$A = QDP, \tag{18}$$

où Q et P sont des matrices orthogonales et D une matrice de la forme (17). Cette décomposition porte le nom de *décomposition singulière* d'une matrice.

Le théorème 1 est complété par le

THÉORÈME 2. *Supposons que la matrice A de l'application* $\mathbf{A}$ *par rapport à un couple de bases orthonormées soit représentée sous forme de produit* (18) *et que les nombres α_i sur la diagonale de la matrice D se disposent dans l'ordre des valeurs décroissantes. Les colonnes des matrices ${}^tP$ et Q sont alors les colonnes de coordonnées des vecteurs de la première et de la seconde base singulière de* $\mathbf{A}$ *respectivement. Les r premiers nombres singuliers de* $\mathbf{A}$ *sont égaux à α_i, $i = 1, \ldots, r$, et les $n - r$ autres sont nuls.*

DÉMONSTRATION. Considérons une colonne p_i de la matrice ${}^tP$. La matrice P étant orthogonale, il vient

$${}^tAAp_i = {}^tP{}^tDDPp_i = {}^tP{}^tDD\boldsymbol{\varepsilon}_i,$$

où $\boldsymbol{\varepsilon}_i$ est une colonne de la matrice unité d'ordre n. La matrice ${}^tDD$ est une matrice diagonale carrée d'ordre n dont les r premiers éléments diagonaux sont α_i^2, et les autres sont nuls. Il en ressort que pour $i \leqslant r$ on a

$${}^tAAp_i = \alpha_i^2\,{}^tP\boldsymbol{\varepsilon}_i = \alpha_i^2 p_i,$$

et ${}^tAAp_i = \mathbf{0}$ pour $i > r$. Donc, p_i, $i = 1, \ldots, n$, sont les matrices-colonnes de coordonnées des vecteurs propres de la transformation $\mathbf{A}^* \mathbf{A}$. On a ainsi

démontré l'assertion concernant la première base singulière et les nombres singuliers.

Ensuite,

$$Ap_i = QDPp_i = QD\varepsilon_i.$$

Soit $\tilde{\varepsilon}_i$ une colonne de la matrice unité d'ordre m. Le produit $D\varepsilon_1$ est égal à $\alpha_i\tilde{\varepsilon}_i$ pour $i \leqslant r$, et à zéro dans les autres cas. On a donc pour tous les $i = 1, \ldots, r$

$$Ap_i = \alpha_i Q\tilde{\varepsilon}_i = \alpha_i q_i,$$

où q_i est une colonne de la matrice Q, ce qui achève la démonstration du théorème.

Si la matrice de l'application **A** par rapport à un couple de bases orthonormées est représentée sous la forme (18), la matrice de l'application adjointe **A*** est

$${}^tA = {}^tP\,{}^tD\,{}^tQ,$$

ce qui entraîne la

PROPOSITION 14. *Soient $e = \|e_1, \ldots, e_n\|$ et $f = \|f_1, \ldots, f_m\|$ la première et la seconde base singulière de l'application* **A**. *Alors f est la première et e la seconde base singulière de l'application* **A***. *Les nombres singuliers non nuls des applications* **A** *et* **A*** *coïncident.*

Admettons que l'application **A** est inversible et que sa matrice par rapport à un couple de bases orthonormées se décompose en produit (18). Alors la matrice de l'application $\mathbf{A}^{-1}$ est

$$A^{-1} = P^{-1}D^{-1}Q^{-1} = {}^tP(D^{-1})\,{}^tQ,$$

et l'on aboutit à la proposition suivante.

PROPOSITION 15. *Supposons que l'application* **A** *est inversible et que e et f sont ses première et seconde bases singulières. Dans ce cas, les bases e et f diffèrent respectivement de la seconde et de la première base singulière de l'application $\mathbf{A}^{-1}$ par l'ordre des vecteurs au plus. Si α_i, $i = 1, \ldots, n$, sont des nombres singuliers de* **A**, *α_i^{-1} sont des nombres singuliers de $\mathbf{A}^{-1}$.*

8. Généralisation aux espaces complexes. Dans le cas d'espaces complexes, on peut définir une application adjointe de la même façon qu'on l'a fait au point 2, et concerver par là même toutes les propriétés de cette application énoncées aux points 2 à 5. Cela est dû au fait qu'on n'y a pas exigé que l'espace vectoriel soit réel. Toutefois la relation qui existe entre l'application adjointe et le produit scalaire dans les espaces euclidiens ne se généralise pas directement aux espaces unitaires. (Cette relation peut être reportée aux espaces euclidiens complexes, mais leur intérêt est minime.)

Pour que la relation entre l'application adjointe et le produit scalaire se maintienne pour les espaces unitaires, il faut modifier la définition de l'espace dual, à quoi on va justement procéder. Considérons un espace vectoriel complexe $\mathscr{L}_n$ et introduisons la définition suivante.

DÉFINITION. La fonction sur $\mathscr{L}_n$ est dite *semi-linéaire* (ou *hermitienne linéaire*) si pour tous $x, y \in \mathscr{L}_n$ elle satisfait aux conditions

$$\langle f, x + y \rangle = \langle f, x \rangle + \langle f, y \rangle ,$$
$$\langle f, \alpha x \rangle = \bar{\alpha} \langle f, x \rangle ,$$

où α est le nombre complexe conjugué de α, et $\langle f, x \rangle$ la valeur de f sur le vecteur x.

Cette définition permet d'exprimer $\langle f, x \rangle$ au moyen des composantes de x dans une base $\boldsymbol{e} = \| e_1, \dots, e_n \|$:

$$\langle f, x \rangle = \sum_i x_i \bar{\xi}^i = \boldsymbol{x}\bar{\boldsymbol{\xi}},$$

où $\bar{\boldsymbol{\xi}}$ est la matrice-colonne conjuguée de la colonne de coordonnées du vecteur x, et $\boldsymbol{x}$ la matrice-ligne des coefficients de la fonction f, avec, comme dans le cas d'une fonction linéaire, $x_i = \langle f, e_i \rangle$.

Définissons, comme pour les fonctions linéaires, l'addition de deux fonctions semi-linéaires et la multiplication d'une fonction semi-linéaire par un nombre au moyen des égalités suivantes :

$$\langle f_1 + f_2, x \rangle = \langle f_1, x \rangle + \langle f_2, x \rangle ,$$
$$\langle \alpha f, x \rangle = \alpha \langle f, x \rangle .$$

Il est facile de vérifier qu'en vertu d'une telle définition d'opérations linéaires, l'ensemble de toutes les fonctions semi-linéaires sur un espace vectoriel complexe $\mathscr{L}_n$ de dimension n est aussi un espace vectoriel complexe de dimension n. C'est cet espace qui sera appelé *dual* ou *dual hermitien* de l'espace $\mathscr{L}_n$ et noté $\mathscr{L}_n^*$.

La base $\| p^1, \dots, p^n \|$ de l'espace $\mathscr{L}_n^*$ est dite *biorthogonale* à la base $\| e_1, \dots, e_n \|$ de l'espace $\mathscr{L}_n$ si

$$\langle p^i, e_j \rangle = \begin{cases} 0, & i \neq j, \\ 1, & i = j. \end{cases}$$

Si $\boldsymbol{x}$ est la matrice-ligne des coefficients de la fonction semi-linéaire f dans la base $\boldsymbol{e}$, la colonne de coordonnées de f dans la base biorthogonale sera ${}^t\boldsymbol{x}$.

Considérons deux espaces vectoriels complexes $\mathscr{L}_n$ et $\bar{\mathscr{L}}_m$ et une application $\mathbf{A} : \mathscr{L}_n \to \bar{\mathscr{L}}_m$. On dit que l'application $\mathbf{A}^* : \bar{\mathscr{L}}_m^* \to \mathscr{L}_n^*$ est *adjointe* de

l'application **A** si pour tous $x \in \mathcal{L}_n$ et $f \in \bar{\mathcal{L}}_m^*$ on a

$$\langle \mathbf{A}^*(f), x \rangle = \langle f, \mathbf{A}(x) \rangle . \tag{19}$$

Ainsi donc, cette définition de l'application adjointe ne diffère pas, par la forme, de la définition correspondante donnée pour les espaces réels. Aussi toutes les propriétés exposées au début de ce paragraphe se conservent-elles avec d'infimes modifications. Notamment, la proposition 2 devient :

PROPOSITION 16. Si *A est une matrice de l'application* $\mathbf{A} : \mathcal{L}_n \to \bar{\mathcal{L}}_m$ *par rapport à un couple de bases,* $A^* = {}^t\bar{A}$ *est celle de l'application adjointe par rapport aux bases biorthogonales correspondantes.*

En effet, choisissons dans $\bar{\mathcal{L}}_m^*$ et $\mathcal{L}_n^*$ deux bases et désignons par $\boldsymbol{\mu}$ et $\boldsymbol{x}$ les matrices-lignes de coordonnées des fonctions semi-linéaires f et $\mathbf{A}^*(f)$ par rapport à ces bases. L'égalité (19) est alors équivalente à l'égalité matricielle $\boldsymbol{\mu}\bar{\boldsymbol{\xi}} = \boldsymbol{x}\overline{A\boldsymbol{\xi}} = \boldsymbol{x}\bar{A}\bar{\boldsymbol{\xi}}$ qui est vérifiée pour toute matrice-colonne $\boldsymbol{\xi}$. Il s'ensuit alors que

$$\boldsymbol{\mu} = \boldsymbol{x}\bar{A}, \tag{20}$$

d'où la proposition.

La proposition 6 se voit également modifiée lorsqu'on l'applique aux transformations linéaires d'espaces complexes.

Dans la proposition 3, toutes les assertions demeurent vraies à l'exception de la formule (8) qui prend la forme

$$(\alpha\mathbf{A})^* = \bar{\alpha}\mathbf{A}^*. \tag{21}$$

La démonstration de la proposition 7 utilise la coïncidence des polynômes caractéristiques des matrices A et ${}^tA$. Comme

$$\det({}^t\bar{A} - \lambda E) = \overline{\det(A - \bar{\lambda}E)},$$

la proposition 7 devient :

PROPOSITION 17. *Si λ est une valeur propre de la transformation* **A** *de l'espace vectoriel complexe* $\mathcal{L}_n$, $\bar{\lambda}$ *est une valeur propre de la transformation adjointe* $\mathbf{A}^* : \mathcal{L}_n^* \to \mathcal{L}_n^*$, *et* $\bar{\lambda}$ *est de la même multiplicité que* λ.

Dans la proposition 9 appliquée à un espace complexe, il faut exiger que les valeurs propres ne soient pas conjuguées, c'est-à-dire que $\lambda \neq \bar{\mu}$, au lieu de $\lambda \neq \mu$ pour les espaces réels.

Les autres résultats des points 2 à 5 sont reportés aux espaces complexes sans être modifiés si l'on entend par un espace dual l'espace dual hermitien.

Considérons un espace unitaire $\mathcal{U}_n$. Etant donné un vecteur $a \in \mathcal{U}_n$, associons à chaque vecteur $x \in \mathcal{U}_n$ le nombre (a, x). Il est aisé de vérifier qu'on définit ainsi une fonction semi-linéaire sur $\mathcal{U}_n$. Inversement, à cha-

que fonction semi-linéaire f sur $\mathscr{U}_n$ on peut associer un vecteur a tel que

$$\langle f, x \rangle = (a, x)$$

pour tout $x \in \mathscr{U}_n$. La relation qui à chaque fonction semi-linéaire f sur $\mathscr{U}_n$ fait associer son vecteur a est un isomorphisme $\boldsymbol{\Gamma} : \mathscr{U}_n^* \to \mathscr{U}_n$ permettant d'identifier ces espaces. Cette identification est courante.

Soit donnée une application linéaire d'espaces unitaires $\mathbf{A} : \mathscr{U}_n \to \tilde{\mathscr{U}}_m$. Vu que chaque espace est identifié avec son dual, l'application adjointe $\mathbf{A}^*$ de $\mathbf{A}$ applique $\tilde{\mathscr{U}}_m$ dans $\mathscr{U}_n$. Elle satisfait à la condition

$$(\mathbf{A}^*(y), x) = (x, \mathbf{A}(x))$$

pour tous $x \in \mathscr{U}_n$ et $y \in \tilde{\mathscr{U}}_m$. Les matrices des applications $\mathbf{A}$ et $\mathbf{A}^*$ sont liées par la relation

$$^t(A^*)\Gamma = \tilde{\Gamma}\bar{A},$$

où Γ et $\tilde{\Gamma}$ sont les matrices de Gram des bases choisies dans $\mathscr{U}_n$ et $\tilde{\mathscr{U}}_m$. Ainsi, la formule (11) prend pour les espaces unitaires la forme

$$A^* = (\Gamma^{-1})^t\bar{A}\tilde{\Gamma}.$$

Avec les définitions introduites, tout le contenu des points 6 et 7, accompagné d'infimes modifications, est reporté au cas des applications d'espaces unitaires. Par exemple, dans l'énoncé du corollaire de la proposition 11, on parlera non pas d'une forme quadratique mais d'une forme quadratique hermitienne.

§ 2. Transformations linéaires dans un espace euclidien

Dans ce paragraphe, on reporte les résultats du § 1 aux transformations dans un espace euclidien. Toutefois, il s'avère commode de s'arrêter au préalable sur quelques propriétés importantes des transformations auto-adjointes, non rattachées au contenu déjà exposé.

1. Transformations commutables. Les transformations $\mathbf{A}$ et $\mathbf{B}$ sont dites *commutables* si $\mathbf{AB} = \mathbf{BA}$. Les propriétés suivantes de ces transformations sont valables aussi bien dans les espaces réels que complexes. Plus loin, la notation Im $\mathbf{B}$ désigne partout l'ensemble des valeurs de la transformation $\mathbf{B}$, noté auparavant $\mathbf{B}(\mathscr{L})$.

PROPOSITION 1. *Si les transformations linéaires* $\mathbf{A}$ *et* $\mathbf{B}$ *de l'espace vectoriel* $\mathscr{L}_n$ *sont commutables, les sous-espaces* Ker $\mathbf{B}$ *et* Im $\mathbf{B}$ *sont invariants par* $\mathbf{A}$.

En effet, supposons que $x \in$ Im $\mathbf{B}$, c'est-à-dire qu'il existe un $y \in \mathscr{L}_n$ tel que $x = \mathbf{B}(y)$. Alors $\mathbf{A}(x) = \mathbf{AB}(y) = \mathbf{BA}(y) \in$ Im $\mathbf{B}$, de sorte que Im $\mathbf{B}$ est invariant.

Soit ensuite $x \in \operatorname{Ker} \mathbf{B}$, c'est-à-dire $\mathbf{B}(x) = o$. Alors, $\mathbf{B}(\mathbf{A}(x)) = \mathbf{AB}(x) = o$. Cela signifie que $\mathbf{A}(x) \in \operatorname{Ker} \mathbf{B}$, et donc $\operatorname{Ker} \mathbf{B}$ est invariant.

Appelons *sous-espace propre* de la transformation $\mathbf{A}$ un sous-espace formé de tous les vecteurs propres associés à une même valeur propre de $\mathbf{A}$, et du vecteur nul.

PROPOSITION 2. *Si* $\mathbf{AB} = \mathbf{BA}$, *les sous-espaces propres de* $\mathbf{B}$ *sont invariants par* $\mathbf{A}$.

Le sous-espace propre associé à la valeur propre λ est $\operatorname{Ker}(\mathbf{B} - \lambda\mathbf{E})$. Si $\mathbf{AB} = \mathbf{BA}$, on a $\mathbf{A}(\mathbf{B} - \lambda\mathbf{E}) = (\mathbf{B} - \lambda\mathbf{E})\mathbf{A}$, ce qui démontre l'assertion en vertu de la proposition 1.

Les transformations $\mathbf{A}$ et $\mathbf{A} - \lambda\mathbf{E}$ étant commutables, on a le

COROLLAIRE. *Le sous-espace* $\operatorname{Im}(\mathbf{A} - \lambda\mathbf{E})$ *est invariant par la transformation* $\mathbf{A}$.

Remarquons que l'invariance de $\operatorname{Ker}(\mathbf{A} - \lambda\mathbf{E})$ est évidente : c'est un sous-espace propre ou un sous-espace nul.

2. Propriétés d'extrémum des valeurs propres. Considérons une transformation symétrique $\mathbf{A}$ de l'espace euclidien $\mathcal{E}_n$. La fonction

$$\rho(x) = \frac{(\mathbf{A}(x), x)}{(x, x)},$$

définie sur tout l'espace $\mathcal{E}_n$ à l'exception du vecteur nul, est appelée *quotient de Rayleigh* de la transformation $\mathbf{A}$ ou de la forme quadratique $(\mathbf{A}(x), x)$. On appellera *sphère unité* $\mathcal{S}_n$ dans $\mathcal{E}_n$ l'ensemble des vecteurs de longueur 1.

PROPOSITION 3. *L'ensemble des valeurs du quotient de Rayleigh se confond avec l'ensemble des valeurs de la forme quadratique* $(\mathbf{A}(x), x)$ *sur la sphère unité.*

En effet, quel que soit le vecteur x de $\mathcal{S}_n$ la valeur de la forme quadratique sur x est égale à la valeur du quotient de Rayleigh sur le même vecteur. Inversement, si pour un $x \in \mathcal{E}_n$ non nul on a $\rho(x) = \alpha$, alors $\alpha = (\mathbf{A}(y), y)$, où $y = |x|^{-1}x \in \mathcal{S}_n$.

On aura besoin de connaître les valeurs maximales et minimales du quotient de Rayleigh. En vertu de la proposition 3, on peut à leur place, considérer les valeurs maximales et minimales de $(\mathbf{A}(x), x)$ sur $\mathcal{S}_n$.

Soit $\|e_1, \ldots, e_n\|$ une base orthonormée des vecteurs propres de la transformation $\mathbf{A}$. On supposera que les vecteurs de base sont ordonnés de manière que les valeurs propres correspondantes forment une suite décroissante : $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$.

Ecrivons la valeur de la forme quadratique sur le vecteur $x \in \mathcal{S}_n$ dans la

base $e = \|e_1, \ldots, e_n\|$:

$$(\mathbf{A}(x), x) = \sum_{i=1}^{n} \lambda_i (\xi^i)^2. \qquad (1)$$

$\sum (\xi^i)^2$ est ici égale à 1 car $x \in \mathcal{S}_n$. Si l'on remplace toutes les valeurs propres par leur valeur maximale, on perturbera en général l'égalité (1) en augmentant son second membre. Donc,

$$\rho(x) = (\mathbf{A}(x), x) \leqslant \lambda_1 \sum_{i=1}^{n} (\xi^i)^2 = \lambda_1. \qquad (2)$$

La borne supérieure mentionnée est atteinte sur le premier vecteur de base :

$$\rho(e_1) = \lambda_1.$$

De façon analogue on démontre que $\rho(x) \geqslant \lambda_n$ et $\rho(e_n) = \lambda_n$. On a ainsi obtenu la

PROPOSITION 4. *La valeur maximale du quotient de Rayleigh est égale à la plus grande valeur propre de la transformation* **A**, *tandis que la valeur minimale, à la plus petite valeur propre. Ces valeurs se présentent sur les vecteurs propres correspondants.*

COROLLAIRE. *Si la forme quadratique* **l** *est définie positive, le rapport des valeurs de deux formes quadratiques* $\mathbf{k}(x)/\mathbf{l}(x)$ *est encadré, pour tout vecteur* x, *par la plus grande et la plus petite racine de l'équation* $\det (A - \lambda B) = 0$, *où* A *et* B *sont les matrices associées aux formes quadratiques* **k** *et* **l** *dans une base* e.

En effet, soient $\boldsymbol{\xi}$ et $\boldsymbol{\eta}$ les colonnes de coordonnées du vecteur x respectivement dans la base e et dans la base canonique de la forme quadratique **l**. Alors, si $\boldsymbol{\xi} = S\boldsymbol{\eta}$, on a ${}^tSBS = E$ et $B = ({}^tS)^{-1}S^{-1}$. Pour le rapport $\mathbf{k}(x)/\mathbf{l}(x)$, on trouve

$$\frac{{}^t\boldsymbol{\xi}A\boldsymbol{\xi}}{{}^t\boldsymbol{\xi}B\boldsymbol{\xi}} = \frac{{}^t\boldsymbol{\eta}\,{}^tSAS\boldsymbol{\eta}}{{}^t\boldsymbol{\eta}\boldsymbol{\eta}}.$$

Introduisons le produit scalaire à l'aide de la forme quadratique **l** (voir théorème 2, § 3, ch. VIII). La base canonique devient alors orthonormée relativement à ce produit scalaire, quant au rapport considéré, il s'avère égal au quotient de Rayleigh de la transformation dont la matrice dans cette base est ${}^tSAS$. Il s'ensuit que $\mathbf{k}(x)/\mathbf{l}(x)$ se trouve encadré par la plus grande et la plus petite racine de l'équation $\det ({}^tSAS - \lambda E) = 0$. Or

$$\det ({}^tSAS - \lambda E) = \det {}^tS \det (A - \lambda ({}^tS)^{-1}S^{-1}) \det S$$

et les racines de l'équation coïncident avec celles de l'équation $\det(A - \lambda B) = 0$. La démonstration est ainsi achevée.

Considérons maintenant un sous-espace $\mathscr{E}_k'$ dans $\mathscr{E}_n$ et la restriction de la forme quadratique $\mathbf{k}(x) = (\mathbf{A}(x), x)$ à $\mathscr{E}_k'$. La restriction de la forme quadratique est une fonction $\mathbf{k}'$ sur $\mathscr{E}_k'$ telle que $\mathbf{k}'(x) = \mathbf{k}(x)$ pour tout x de $\mathscr{E}_k'$. Il est évident que $\mathbf{k}'$ est une forme quadratique sur $\mathscr{E}_k'$ et qu'elle possède une base orthonormée dans $\mathscr{E}_k'$ par rapport à laquelle elle se décompose en carrés :

$$\mathbf{k}'(x) = \sum_{i=1}^{k} \mu_i (\xi^i)^2.$$

Admettons que les vecteurs de base sont numérotés de manière que $\mu_1, \ldots$ $\ldots, \mu_k$ forment une suite décroissante : $\mu_1 \geqslant \mu_2 \geqslant \ldots \geqslant \mu_k$. Dans ce cas, μ_1 est le maximum du quotient de Rayleigh pour $\mathbf{k}'(x)$ et, partant, la valeur maximale qu'il prend pour $\mathbf{k}(x)$ sur le sous-espace $\mathscr{E}_k'$.

Proposition 5. *Soient $\lambda_1, \ldots, \lambda_n$ les valeurs propres de la transformation symétrique $\mathbf{A}$, et $\mu_1, \ldots, \mu_k$ les valeurs propres de sa restriction au sous-espace invariant k-dimensionnel, numérotées dans l'ordre décroissant. Alors*

$$\lambda_1 \geqslant \mu_1 \geqslant \lambda_{n-k+1}. \tag{3}$$

En effet, la sphère unité $\mathscr{S}_k$ de l'espace $\mathscr{E}_k'$ est contenue dans la sphère unité de l'espace $\mathscr{E}_n$, et la valeur maximale de la fonction sur un sous-ensemble est au plus égale à sa valeur maximale sur tout l'ensemble, de sorte que $\lambda_1 \geqslant \mu_1$.

Démontrons la seconde inégalité. A cet effet, choisissons dans $\mathscr{E}_k'$ un vecteur x dont les coordonnées dans la base canonique de la forme quadratique $\mathbf{k}(x)$ sont $(\xi^1, \ldots, \xi^{n-k+1}, 0, \ldots, 0)$, c'est-à-dire un vecteur situé dans l'intersection de $\mathscr{E}_k'$ avec l'enveloppe linéaire des $n - k + 1$ premiers vecteurs de base. L'intersection de ces espaces contient un vecteur non nul car la somme de leurs dimensions est strictement supérieure à n. Soit $y = |x|^{-1}x$. En substituant λ_{n-k+1} à toutes les valeurs propres, on obtient

$$\mathbf{k}'(y) = \mathbf{k}(y) = \lambda_1(\eta^1)^2 + \ldots + \lambda_{n-k+1}(\eta^{n-k+1})^2 \geqslant \lambda_{n-k+1},$$

d'où l'inégalité nécessaire.

Pour un sous-espace donné $\mathscr{E}_k'$, la quantité μ_1 est une constante, mais qui varie avec le passage à un autre sous-espace $\mathscr{E}_k''$ tout en vérifiant la double inégalité (3). Ceci étant, elle peut atteindre soit λ_1, soit λ_{n-k+1}. En effet, désignons par $\mathscr{E}_k$ l'enveloppe linéaire des vecteurs $e_1, \ldots, e_k$. Dans ce sous-espace, le quotient de Rayleigh atteint sur le vecteur e_1 son maximum absolu, soit : $\lambda_1 = \rho(e_1)$, de sorte que $\mu_1 = \lambda_1$. Pour l'enveloppe linéaire

des vecteurs $e_{n-k+1}, \ldots, e_n$, on a $\rho(x) = \lambda_{n-k+1}(\eta^{n-k+1})^2 + \ldots$ $\ldots + \lambda_n(\eta^n)^2$, où $\sum_{i=n-k+1}^{n} (\eta^i)^2 = 1$. Donc $\rho(x) \leqslant \lambda_{n-k+1}$ et $\mu_1 =$ $= \rho(e_{n-k+1}) = \lambda_{n-k+1}$.

On arrive ainsi au théorème suivant.

THÉORÈME 1. *Etant donné que les valeurs propres d'une transformation linéaire symétrique forment une suite décroissante* $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$, *on a*

$$\lambda_{n-k+1} = \min_{\mathscr{E}_k} \max_{0 \neq x \in \mathscr{E}_k} \rho(x),$$

où le maximum est pris suivant tous les vecteurs non nuls du sous-espace $\mathscr{E}_k$, *et le minimum suivant tous les sous-espaces invariants de dimension k.*

On a vu en effet que λ_{n-k+1} est la borne inférieure atteinte pour μ_1 et que μ_1 est égal à $\max\limits_{0 \neq x \in \mathscr{E}_k} \rho(x)$ selon la proposition 4.

Notons $\mathscr{E}'_{n-k}$ l'enveloppe linéaire des $n - k$ premiers vecteurs de base $e_1, \ldots, e_{n-k}$. Son supplémentaire orthogonal $\mathscr{E}'_{n-k}$ contient le vecteur e_{n-k+1} pour lequel $\rho(e_{n-k+1}) = \lambda_{n-k+1}$. Cela signifie que le minimum dont il s'agit dans le théorème 1 est atteint pour le sous-espace $(\mathscr{E}'_{n-k})^\perp$. Ces considérations permettent de se représenter comment se construit la base orthonormée formée des vecteurs propres de la transformation **A**, processus qu'on décrira sans entrer dans les détails de sa réalisation.

On recherche d'abord le vecteur sur lequel est atteint le maximum du quotient de Rayleigh, plus précisément l'un de ces vecteurs. Ensuite, dans le supplémentaire orthogonal du sous-espace engendré par ce vecteur, on cherche de nouveau le vecteur sur lequel est atteint le maximum. Après quoi, on recherche le maximum dans le supplémentaire orthogonal de l'enveloppe linéaire engendrée par deux vecteurs choisis, etc. Il ne reste finalement qu'à normer le système des vecteurs obtenus.

Pour justifier cette construction, il faut démontrer que le maximum du quotient de Rayleigh n'est atteint que sur le vecteur propre. On le réalise en recourant à la méthode de Lagrange permettant de trouver l'extrémum lié (voir Koudriavtsev [21], t. II, § 42). La fonction de Lagrange est de la forme

$$\sum_{i,j=1}^{n} a_{ij}\xi^i\xi^j - \lambda \sum_{i=1}^{n} (\xi^i)^2.$$

En égalant à zéro ses dérivées partielles, on obtient le système d'équations

que vérifient les vecteurs propres :

$$a_{11}\xi^1 + \ldots + a_{1n}\xi^n - \lambda\xi^1 = 0,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$a_{n1}\xi^1 + \ldots + a_{nn}\xi^n - \lambda\xi^n = 0.$$

3. Décomposition polaire. On sait que la transformation affine du plan se décompose en produit d'une transformation orthogonale et de deux contractions vers des droites perpendiculaires.

On va maintenant généraliser ce résultat pour une transformation linéaire **A** de l'espace euclidien. La décomposition dont il s'agit s'appelle *décomposition polaire* de **A**.

THÉORÈME 2. *Pour toute transformation linéaire* **A** *de l'espace euclidien* $\mathcal{E}_n$ *il existe une transformation symétrique* **S** *à valeurs propres positives et une transformation orthogonale* **P** *telles que*

$$\mathbf{A} = \mathbf{PS}. \tag{4}$$

Avant de démontrer le théorème, il convient de considérer la proposition suivante.

PROPOSITION 6. *Si* $\|e_1, \ldots, e_n\|$ *et* $\|f_1, \ldots, f_n\|$ *sont des bases orthonormées dans* $\mathcal{E}_n$, *il existe une transformation orthogonale* **P** *telle que* $\mathbf{P}(e_i) = f_i$ $(i = 1, \ldots, n)$.

En effet, la condition $\mathbf{P}(e_i) = f_i$ définit de façon univoque toutes les colonnes de la matrice associée à la transformation **P** dans la base e, ce sont les colonnes de coordonnées des vecteurs $f_1, \ldots, f_n$. Notons-les $\boldsymbol{\varphi}_i$. Vu que le système de vecteurs f est orthonormé, il vient

$${}^t\boldsymbol{\varphi}_i\boldsymbol{\varphi}_j = \begin{cases} 0, & i \neq j, \\ 1, & i = j, \end{cases}$$

de sorte que la matrice de la transformation est orthogonale, d'où la conclusion recherchée.

Passons maintenant à la démonstration du théorème 2. Soient $e = \|e_1, \ldots, e_n\|$ et $f = \|f_1, \ldots, f_n\|$ les bases singulières de la transformation **A**. (Comme **A** est une transformation, les deux bases sont des bases dans $\mathcal{E}_n$.) On a alors pour tous les $i = 1, \ldots, n$

$$\mathbf{A}(e_i) = \alpha_i f_i,$$

où α_i sont les nombres singuliers de la transformation **A**. En effet, pour $i \leqslant \operatorname{Rg} \mathbf{A}$, c'est la définition des vecteurs f_i, et pour les autres i on a d'une part $\alpha_i = 0$ et d'autre part $\mathbf{A}(e_i) = o$, si bien que l'égalité est vérifiée.

Considérons maintenant une transformation orthogonale **P** qui à la première base singulière fait correspondre la seconde : $\mathbf{P}(e_i) = f_i$ $(i =$

$= 1, \ldots, n$. Démontrons que la transformation $\mathbf{S} = \mathbf{P}^{-1}\mathbf{A}$ est symétrique. En effet, pour tout i on a

$$\mathbf{P}^{-1}\mathbf{A}(e_i) = \mathbf{P}^{-1}(\alpha_i f_i) = \alpha_i e_i. \tag{5}$$

Ainsi donc, la transformation $\mathbf{S}$ possède une base orthonormée des vecteurs propres $e_1, \ldots, e_n$. Sa matrice y est diagonale et par suite, symétrique. Donc, $\mathbf{S}$ est une transformation symétrique. On voit en outre de (5) que les valeurs propres de $\mathbf{S}$ sont égaux à $\alpha_1, \ldots, \alpha_n$ et par suite, sont positives. On achève ainsi la démonstration du théorème.

Le théorème 2 peut être formulé en termes de matrices.

THÉORÈME 2M. *Pour toute matrice carrée A il existe une matrice orthogonale P et une matrice symétrique S, avec nombres caractéristiques positifs, telles que*

$$A = PS. \tag{6}$$

PROPOSITION 7. *Pour toute transformation linéaire* $\mathbf{A}$ *de l'espace euclidien* $\mathcal{E}_n$ *il existe une transformation orthogonale* $\mathbf{P}'$ *et une transformation symétrique* $\mathbf{S}'$ *à valeurs propres positives, telles que* $\mathbf{A} = \mathbf{S}'\mathbf{P}'$.

Pour le démontrer, écrivons la décomposition (4) pour la transformation $\mathbf{A}^*$. Soit $\mathbf{A}^* = \mathbf{P}_1\mathbf{S}_1$. On a alors $\mathbf{A} = \mathbf{S}_1^*\mathbf{P}_1^* = \mathbf{S}_1\mathbf{P}_1^{-1}$ et l'on peut poser $\mathbf{S}' = \mathbf{S}_1$ et $\mathbf{P}' = \mathbf{P}_1^{-1}$.

4. Unicité de la décomposition polaire. Si la transformation $\mathbf{A}$ est représentée sous forme de $\mathbf{PS}$, où $\mathbf{S}$ est une transformation symétrique et $\mathbf{P}$ une transformation orthogonale, on a $\mathbf{A}^*\mathbf{A} = \mathbf{S}^*\mathbf{P}^*\mathbf{PS} = \mathbf{S}^2$. Montrons que la transformation $\mathbf{S}$ se définit d'après $\mathbf{S}^2$ et partant d'après $\mathbf{A}$ de façon unique. Démontrons d'abord pour cela la proposition suivante.

PROPOSITION 8. *Les vecteurs propres de la transformation* $\mathbf{S}$ *sont des vecteurs propres de la transformation* $\mathbf{S}^2$. *Si* $\mathbf{S}$ *est une transformation symétrique à valeurs propres positives, les vecteurs propres de la transformation* $\mathbf{S}^2$ *sont aussi propres pour* $\mathbf{S}$. *Dans ce cas, les valeurs propres de la transformation* $\mathbf{S}$ *sont égales aux racines carrées arithmétiques des valeurs propres correspondantes de* $\mathbf{S}^2$.

DÉMONSTRATION. 1) Soit $\mathbf{S}(x) = \alpha x$ pour un vecteur $x \neq o$. Alors $\mathbf{S}^2(x) = \mathbf{S}(\mathbf{S}(x)) = \mathbf{S}(\alpha x) = \alpha^2 x$.

2) Soit $\mathbf{S}^2(x) = \lambda x$ et soient ξ^i les coordonnées du vecteur x par rapport à la base orthonormée $e = \| e_1, \ldots, e_n \|$ composée des vecteurs propres de la transformation $\mathbf{S}$. On a alors

$$\mathbf{S}^2(x) = \lambda \sum_i \xi^i e_i = \sum_i \lambda \xi^i e_i.$$

D'autre part, si α_i sont les valeurs propres de **S**, on a

$$\mathbf{S}^2(x) = \sum_i \xi^i \mathbf{S}^2(e_i) = \sum_i \xi^i \alpha_i^2 e_i .$$

Il en découle que $\xi^i(\alpha_i^2 - \lambda) = 0$ pour tous les $i = 1, \ldots, n$. Ainsi, $\xi^i = 0$ pour tout i tel que $\alpha_i^2 \neq \lambda$. Vu que α_i est positif, les égalités $\alpha_i^2 = \lambda$ et $\alpha_j^2 = \lambda$ ne peuvent se vérifier qu'à la condition $\alpha_i = \alpha_j$. Le vecteur x se décompose donc suivant les vecteurs de la base e, qui correspondent à une même valeur propre. Cela signifie qu'il est également un vecteur propre pour la transformation **S** et qu'il est associé à la valeur propre $\alpha_i = \sqrt{\lambda}$. La proposition est démontrée.

L'unicité de la transformation **S** résulte maintenant de la proposition suivante.

PROPOSITION 9. *La transformation symétrique* **S** *se définit de façon univoque par ses valeurs propres et ses vecteurs propres.*

Pour le démontrer, considérons une famille de sous-espaces $\mathscr{L}_1, \ldots, \mathscr{L}_p$ dont chacun est engendré par les vecteurs propres associés à une même valeur propre et par le vecteur nul. Il y a autant de sous-espaces que de valeurs propres différentes. Chaque vecteur se décompose de façon unique en une somme de la forme

$$x = x_1 + \ldots + x_p, \quad \text{où} \quad x_\alpha \in \mathscr{L}_\alpha .$$

En effet, on peut obtenir une telle décomposition en groupant les termes dans la décomposition de x suivant toute base de vecteurs propres. Démontrons l'unicité de la décomposition. Si on en avait deux : $x = x_1 + \ldots + x_p = y_1 + \ldots + y_p$, on aurait $(x_1 - y_1) + \ldots + (x_p - y_p) = o$. La différence $x_\alpha - y_\alpha$ de $\mathscr{L}_\alpha$ appartient également dans ce cas à la somme de tous les autres sous-espaces. Si elle n'est pas nulle, il en découle aussitôt la dépendance linéaire des vecteurs propres associés aux valeurs propres différentes.

Si donc la décomposition de x est définie de façon univoque, il en est de même de son image $\mathbf{S}(x)$:

$$\mathbf{S}(x) = \alpha_1 x_1 + \ldots + \alpha_p x_p .$$

La proposition est démontrée.

On pourrait sans peine exprimer $\mathbf{S}(x)$ dans la base orthonormée des vecteurs propres :

$$\mathbf{S}(x) = \sum_{i=1}^{n} \alpha_i (x, e_i) e_i , \tag{7}$$

mais dans ce cas il aurait fallu montrer que le résultat est indépendant du choix de la base.

Dans le cas général, la transformation orthogonale **P** qui figure dans la décomposition polaire de la transformation **A** n'est pas parfaitement définie. Toutefois, si la transformation **A** n'est pas dégénérée, autrement dit, si Rg **A** $= n$, il en est de même de la transformation **S** et, par suite, **P** est parfaitement défini comme $\mathbf{AS}^{-1}$.

5. Nombres singuliers et bases singulières d'une transformation. On a introduit dans le § 1 les nombres singuliers et les bases singulières d'une application d'un espace euclidien dans l'autre. Maintenant on les étudiera en détail pour le cas d'une transformation.

Il découle de la formule (5) obtenue lors de la démonstration du théorème 2 que les nombres singuliers d'une transformation **A** sont égaux aux valeurs propres de la transformation symétrique **S** qui figure dans la décomposition polaire de **A**, et que la première base singulière de **A** est constituée des vecteurs propres de **S**.

Voyons quelle est l'interprétation géométrique des nombres singuliers dans l'espace euclidien tridimensionnel. L'image par la transformation **A** de la sphère unité $\mathscr{S}_3$ est un ellipsoïde. Considérons un rayon quelconque de cet ellipsoïde, c'est-à-dire la longueur d'un vecteur $\mathbf{A}(x)$ tel que $(x, x) = 1$. Il vient,

$$|\mathbf{A}(x)|^2 = (\mathbf{A}(x), \mathbf{A}(x)) = (\mathbf{A}^*\mathbf{A}(x), x). \tag{8}$$

Pour le vecteur x situé sur la sphère unité, cette expression coïncide avec le quotient de Rayleigh pour la transformation $\mathbf{A}^*\mathbf{A}$ dont les valeurs propres sont les carrés des nombres singuliers. Il ressort donc de la proposition 4 que

$$\alpha_1 \geqslant |\mathbf{A}(x)| \geqslant \alpha_3$$

pour tous les $x \in \mathscr{S}_3$. Cela signifie que α_1 et α_3 sont le plus grand et le plus petit rapport des longueurs de l'image et de son antécédent par la transformation **A**. Il est évident que α_1 et α_3 sont les demi-grand et -petit axes de l'ellipsoïde.

Ensuite, selon le théorème 1,

$$\alpha_2 = \min_{\mathscr{S}_2} \max_{0 \neq x \in \mathscr{S}_2} |\mathbf{A}(x)|.$$

Le minimum est atteint sur le plan orthogonal au grand axe de l'ellipsoïde, avec max $|\mathbf{A}(x)|$ égal au demi-grand axe de l'ellipse située dans ce plan. Par suite, α_2 est le troisième demi-axe de l'ellipsoïde.

Ainsi, les nombres singuliers caractérisent une traction de l'espace engendrée par la transformation **A**, indépendamment de la rotation des vecteurs au cours de la transformation. Pour une transformation symétri-

que ils sont égaux aux modules des valeurs propres, mais ils sont aussi définis dans le cas d'une transformation quelconque. Leur signification géométrique peut être facilement généralisée à un espace de dimension quelconque, car la formule (8) ne dépend pas de la dimension de l'espace. Notons en particulier la proposition suivante.

PROPOSITION 10. *Pour toute transformation linéaire* **A** *de l'espace euclidien* $\mathcal{E}_n$,

$$\max_{o \neq x \in \mathcal{E}_n} \frac{|\mathbf{A}(x)|}{|x|} = \alpha_1, \qquad \min_{o \neq x \in \mathcal{E}_n} \frac{|\mathbf{A}(x)|}{|x|} = \alpha_n,$$

où α_1 *et* α_n *sont le plus grand et le plus petit nombre singulier de la transformation* **A**.

En effet, en posant $y = |x|^{-1}x$, on déduit de la formule (8) que le carré du rapport $|\mathbf{A}(x)|/|x|$ est égal au quotient de Rayleigh pour la transformation $\mathbf{A}^*\mathbf{A}$. D'où en vertu de la proposition 4, on obtient la conclusion recherchée.

Donnons quelques propriétés des nombres singuliers. Le déterminant et la trace de la matrice d'une transformation linéaire ainsi que tous les coefficients de son polynôme caractéristique, sont les invariants de la transformation. Les propositions suivantes indiquent le lien qui les relie aux nombres singuliers.

PROPOSITION 11. *Le déterminant de la matrice de la transformation linéaire dans un espace euclidien est égal en valeur absolue au produit des nombres singuliers.*

Pour le démontrer, calculons le module du déterminant de chacun des deux membres de l'égalité (6). On obtient

$$|\det A| = |\det P| \cdot |\det S|.$$

$|\det P|$ est ici égal à 1 car la transformation **P** est orthogonale, et $\det S$ vaut le produit des valeurs propres de la transformation **S**, égales aux nombres singuliers. D'où $|\det A| = \alpha_1 \ldots \alpha_n$, ce qu'il fallait démontrer.

PROPOSITION 12. *La valeur absolue de la trace de la matrice d'une transformation linéaire* **A** *ne dépasse pas la somme de ses nombres singuliers.*

Pour le démontrer, écrivons la décomposition polaire de **A** dans la première base singulière et représentons sa matrice sous la forme $A = PS$, où S est la matrice diagonale avec nombres singuliers sur la diagonale. Il est aisé de vérifier que lorsqu'on multiplie P par la matrice diagonale, chaque colonne de P devient multipliée par un élément correspondant de la diagonale de S. Aussi la trace de A est-elle égale à la somme des produits des éléments diagonaux correspondants de P et S. Or les éléments de la matrice

orthogonale sont inférieurs en valeur absolue à l'unité. D'où

$$|\operatorname{tr} A| \leqslant \sum |p_{ii}\alpha_i| \leqslant \sum \alpha_i,$$

ce qu'il fallait démontrer.

PROPOSITION 13. *Les nombres singuliers de la transformation* **A** *ne changent pas quand on la multiplie à gauche ou à droite par la transformation orthogonale* **U**.

DÉMONSTRATION. Soit A la matrice de la transformation **A** dans une base orthonormée. Ecrivons sa décomposition singulière (18) du § 1 : $A = QDP$, où Q et P sont des matrices orthogonales, et D une matrice diagonale carrée d'ordre n. Les nombres singuliers de la transformation **A** sont les éléments diagonaux de la matrice D. La matrice de la transformation **AU** est $AU = QDPU$, où PU est une matrice orthogonale. En appliquant le théorème 2 du § 1 on constate que les éléments diagonaux de la matrice D sont les nombres singuliers de la transformation **AU**.

Pour le produit **UA**, la proposition se démontre de façon analogue.

PROPOSITION 14. *Les nombres singuliers de la transformation* **A** *coïncident avec les nombres singuliers de la transformation adjointe* **A***.

Cette proposition découle directement de la proposition 14 du § 1.

PROPOSITION 15. *L'image de tout vecteur par une transformation peut être obtenue si sont connus les nombres singuliers et les bases singulières de cette transformation.*

En effet, écrivons le développement d'un vecteur quelconque dans la première base singulière :

$$x = \sum_i (x, e_i)e_i$$

et considérons les images des deux membres de l'égalité par la transformation symétrique **S** intervenant dans la décomposition polaire **A** = **PS** :

$$\mathbf{S}(x) = \sum_i (x, e_i)\mathbf{S}(e_i) = \sum_i \alpha_i(x, e_i)e_i.$$

Recherchons ensuite les images des deux membres de cette égalité par la transformation orthogonale **P**. On obtient la décomposition

$$\mathbf{A}(x) = \sum_i \alpha_i(x, e_i)\mathbf{P}(e_i) = \sum_i \alpha_i(x, e_i)f_i,$$

où f_i sont les vecteurs de la seconde base singulière, et α_i les nombres singuliers.

A titre d'exercice, le lecteur peut essayer d'obtenir la proposition 15 à partir du théorème 2, § 1, et les propositions 13 et 14 s'il utilise la décomposition polaire.

6. Etude des résultats pour les espaces unitaires. Considérons une transformation auto-adjointe $\mathbf{A}$ de l'espace unitaire $\mathscr{U}_n$. Le quotient de Rayleigh se définit pour $\mathbf{A}$ de la même façon que dans le cas d'un espace euclidien. Etant donné la base orthonormée des vecteurs propres de la transformation $\mathbf{A}$, on peut écrire

$$\rho(x) = \frac{(\mathbf{A}(x), x)}{(x, x)} = \frac{\sum \lambda_i \xi^i \xi^i}{\sum \xi_i \xi^i},$$

où λ_i sont les valeurs propres de la transformation $\mathbf{A}$. Il s'ensuit une assertion équivalente à la proposition 4 : pour tout x, on a

$$\lambda_n \leqslant \rho(x) \leqslant \lambda_1 \tag{9}$$

(à condition que les valeurs propres forment une suite décroissante).

Les propriétés de l'extrémum des valeurs propres, qui ont été formulées au théorème 1, sont reportées sans changements aux transformations auto-adjointes dans les espaces unitaires.

La décomposition polaire de la transformation linéaire dans un espace unitaire est de la forme

$$\mathbf{A} = \mathbf{PS},$$

où $\mathbf{P}$ est une transformation unitaire et $\mathbf{S}$ une transformation auto-adjointe à valeurs propres positives. La démonstration de l'existence et de l'unicité de cette décomposition pour des transformations inversibles ne diffère pratiquement pas de la démonstration donnée pour les espaces euclidiens : on prend pour $\mathbf{P}$ la transformation unitaire qui à la première base singulière fait correspondre la seconde, tandis que $\mathbf{S} = \mathbf{P}^{-1}\mathbf{A}$ est la transformation auto-adjointe dont les valeurs propres sont égales aux nombres singuliers de $\mathbf{A}$.

Les propriétés des nombres singuliers, obtenues au point 5, sont reportées presque sans changement aux espaces unitaires. On a ainsi la

PROPOSITION 16. *Les nombres singuliers de la transformation* $\mathbf{A}$ *ne changent pas quand on la multiplie à droite ou à gauche par une transformation unitaire* $\mathbf{Q}$.

7. Réduction de la matrice d'une transformation linéaire à la forme triangulaire.

PROPOSITION 17. *Chaque transformation linéaire* $\mathbf{A}$ *d'un espace vectoriel complexe* $\mathscr{L}_n$ *présente un vecteur propre. Tout sous-espace invariant par la transformation* $\mathbf{A}$ *contient un vecteur propre.*

La première assertion est évidente, vu que l'équation caractéristique a au moins une racine qui est une valeur propre. La deuxième assertion découle de la première si celle-ci est appliquée à la restriction de la transformation considérée sur un sous-espace invariant.

Les propositions 4 et 17 entraînent la

PROPOSITION 18. *Tout couple de transformations commutables dans un espace complexe possède un vecteur propre commun.*

En effet, soient **A** et **B** deux transformations qui commutent. Tout espace propre de la transformation **B** est invariant par **A** et par suite, contient un vecteur propre de **A**.

La proposition 17 signifie précisément que tout sous-espace invariant contient un sous-espace invariant unidimensionnel. On peut de même démontrer la proposition suivante.

PROPOSITION 19. *Tout espace k-dimensionnel invariant par la transformation linéaire* **A** *dans l'espace complexe* $\mathscr{L}_n$, *contient un sous-espace invariant* $(k - 1)$*-dimensionnel.*

Démontrons d'abord que pour la transformation **A** dans l'espace vectoriel complexe $\mathscr{L}_n$ il existe un sous-espace $(n - 1)$-dimensionnel $\mathscr{L}_{n-1}$ qui est invariant par **A**. Pour le faire, considérons le sous-espace Im $(\mathbf{A} - \lambda\mathbf{E})$, où λ est une valeur propre de **A**. Ce sous-espace est, comme on l'a vu, invariant par **A** et est de dimension Rg $(\mathbf{A} - \lambda\mathbf{E})$ strictement inférieure à n. Elle peut même s'avérer strictement inférieure à $n - 1$. Dans ce cas, démontrons que tout sous-espace $\mathscr{L}_{n-1}$ comprenant Im $(\mathbf{A} - \lambda\mathbf{E})$ est invariant par **A**.

En effet, posons Im $(\mathbf{A} - \lambda\mathbf{E}) \subsetneq \mathscr{L}_{n-1}$. Alors $(\mathbf{A} - \lambda\mathbf{E})x \in \mathscr{L}_{n-1}$ pour tout vecteur x. Maintenant si $x \in \mathscr{L}_{n-1}$, il découle alors de $\mathbf{A}(x) - \lambda x \in \in \mathscr{L}_{n-1}$ que $\mathbf{A}(x) \in \mathscr{L}_{n-1}$, ce qui signifie justement que le sous-espace $\mathscr{L}_{n-1}$ est invariant.

Pour démontrer la proposition pour un k quelconque, il suffit maintenant d'appliquer l'assertion démontrée à la restriction de la transformation **A** sur un sous-espace invariant k-dimensionnel.

Il découle de la proposition 19 déjà démontrée que pour toute transformation linéaire d'un espace vectoriel complexe $\mathscr{L}_n$ il existe une suite de sous-espaces invariants injectés l'un dans l'autre $\mathscr{L}_{n-1}, \mathscr{L}_{n-2}, \ldots, \mathscr{L}_1$ de dimensions $n - 1, n - 2, \ldots, 1$:

$$\mathscr{L}_1 \subset \mathscr{L}_2 \subset \ldots \subset \mathscr{L}_{n-2} \subset \mathscr{L}_{n-1} \subset \mathscr{L}_n. \tag{10}$$

En effet, pour construire ces sous-espaces, on peut reprendre plusieurs fois la proposition (19) : à chaque étape, on l'applique à un sous-espace invariant construit à l'étape précédente.

DÉFINITION. On dit que la matrice A d'éléments a^i_j est *triangulaire supérieure* (resp. *inférieure*) si $a^i_j = 0$ pour $i > j$ (resp. pour $i < j$).

PROPOSITION 20. *Pour toute transformation linéaire* $\mathbf{A}$ *de l'espace vectoriel complexe* $\mathscr{L}_n$ *il existe une base dans laquelle sa matrice est triangulaire supérieure.*

Pour le démontrer, construisons une base en nous appuyant sur les sous-espaces invariants (10) de la façon suivante : $e_1 \in \mathscr{L}_1$, le vecteur e_2 forme avec e_1 une base dans $\mathscr{L}_2$, le vecteur e_3 complète e_1, e_2 jusqu'à une base de $\mathscr{L}_3$, etc. Ainsi, pour tous les $k = 1, \ldots, n$, les vecteurs $e_1, \ldots, e_k$ constituent une base dans $\mathscr{L}_k$. Vu que les sous-espaces $\mathscr{L}_k$ sont invariants, le vecteur $\mathbf{A}(e_k)$ appartient au sous-espace $\mathscr{L}_k$ pour tout k et, partant, se développe suivant les vecteurs $e_1, \ldots, e_k$. Il s'ensuit que pour les éléments a^i_k de la matrice de $\mathbf{A}$ on a $a^i_k = 0$ pour $i > k$, ce qu'il fallait montrer.

Pour les espaces unitaires la proposition 20 devient :

THÉORÈME 3. *Pour toute transformation linéaire d'un espace unitaire* $\mathscr{U}_n$ *il existe une base orthonormée dans laquelle sa matrice est triangulaire supérieure.*

Pour le démontrer, il suffit de noter que la base construite dans la proposition 20 est dans une large mesure arbitraire et elle peut être choisie orthonormée. En effet, en utilisant le procédé d'orthogonalisation de Gram-Schmidt, on ajoute à tout vecteur e_k un vecteur qui appartient à l'enveloppe linéaire engendrée par $e_1, \ldots, e_{k-1}$, c'est-à-dire à $\mathscr{L}_{k-1}$, de sorte que le vecteur qui en résulte appartient encore à $\mathscr{L}_k$.

§ 3. Espaces normés

1. Définition. Dans nombre de problèmes liés aux espaces vectoriels, il s'avère nécessaire de comparer deux éléments d'un espace, par exemple de pouvoir affirmer qu'un vecteur est dans un certain sens plus petit qu'un autre. Dans un espace euclidien, il est naturel de comparer les vecteurs en longueur. Dans les autres espaces, on peut aussi introduire un produit scalaire, mais souvent la nature de leurs éléments est telle qu'il n'existe aucun produit scalaire qui lui soit naturellement lié. Par ailleurs, il se peut que le produit scalaire ne soit pas nécessaire et on n'a besoin que d'un certain analogue de la longueur du vecteur, à savoir d'une fonction numérique de vecteur présentant quelques propriétés importantes. Ces fonctions sont introduites au moyen de la définition suivante.

DÉFINITION. Soit une fonction φ qui à chaque vecteur d'un espace vectoriel réel ou complexe $\mathscr{L}$ associe un nombre réel. Cette fonction est appelée *norme* et sa valeur sur le vecteur x, norme de ce vecteur, si sont satisfaites

les conditions suivantes pour tous vecteurs x et y et tout nombre λ :

1) $\varphi(x) > 0$ pour tout $x \neq o$;

2) $\varphi(\lambda x) = |\lambda|\varphi(x)$ (« homogénéité positive ») ;

3) $\varphi(x + y) \leqslant \varphi(x) + \varphi(y)$ (« convexité »).

L'espace vectoriel sur lequel est définie une norme est dit *normé*. On note souvent $\|x\|$ la norme du vecteur x. A la place de l'expression « norme dont la valeur sur le vecteur x est noté $\|x\|$ », on dira et écrira « norme $\| * \|$ ».

Remarquons que la propriété 2) entraîne $\varphi(o) = 0$.

Une autre propriété des normes s'écrit par l'inégalité

$$\varphi(x - y) \geqslant |\varphi(x) - \varphi(y)|. \tag{1}$$

En effet,

$$\varphi(x) - \varphi(y) = \varphi(x - y + y) - \varphi(y) \leqslant$$
$$\leqslant \varphi(x - y) + \varphi(y) - \varphi(y) = \varphi(x - y).$$

De façon analogue on démontre que

$$\varphi(y) - \varphi(x) \leqslant \varphi(y - x) = \varphi(x - y).$$

Dans un espace normé, on est en mesure de définir la *distance* entre les vecteurs x et y comme la norme de leur différence $\|y - x\|$. La distance ainsi définie possède les propriétés caractéristiques de la distance entre les points de l'espace euclidien : elle est positive et s'annule si et seulement si x coïncide avec y. En outre, $\|y - x\| = \|x - y\|$, c'est-à-dire que la distance est symétrique, et l'inégalité triangulaire est vérifiée : $\|x - y\| + \|y - z\| \geqslant \|x - z\|$.

L'ensemble des vecteurs d'un espace normé, dont la distance à un vecteur a ne dépasse pas un nombre donné ε est appelé *ε-voisinage* du vecteur a.

En utilisant la notion de voisinage, on est en mesure de définir la limite d'une suite de vecteurs dans l'espace normé : le vecteur a est appelé *limite* de la suite de vecteurs $\{x_k\}$ si pour chaque $\varepsilon > 0$ il existe un nombre $k_0(\varepsilon)$ tel que tous les éléments de la suite à partir de l'élément de numéro k_0 se trouvent dans le ε-voisinage du vecteur a, ou plus brièvement, chaque voisinage du vecteur a contient tous les éléments de la suite à l'exception d'un ensemble fini.

Ainsi, il apparaît la possibilité de reporter sur les espace normés, sous une forme ou l'autre, toutes les notions d'analyse mathématique élémentaire. Notre objectif n'est pas d'exposer cette théorie. On peut s'initier à ses éléments en s'adressant aux cours d'analyse mathématique.

Les espaces normés jouent en analyse mathématique un rôle très important. Cela s'explique par le fait que les problèmes d'analyse mathématique

se rapportent le plus souvent non pas à des fonctions isolées, mais à des classes de fonctions. Ces classes ont des structures d'espaces vectoriels de dimension infinie, normés ou même plus généraux, dits espaces topologiques vectoriels. Le domaine des mathématiques étudiant ces espaces est appelé analyse fonctionnelle.

En algèbre linéaire, on étudie les espaces vectoriels de dimension finie, de sorte que les notions d'analyse se rattachant au passage à la limite ne joueront pas pour nous un rôle prépondérant. On discutera dans ce paragraphe des normes les plus courantes sur les espaces vectoriels de dimension finie et en particulier, des normes sur les espaces des matrices. On montrera toutefois plus loin que sous l'angle de convergeance des suites en norme la différence entre les normes sur un espace de dimension finie est de peu d'importance.

2. Exemples de normes. Considérons un espace arithmétique de dimension n, réel ou complexe, c'est-à-dire l'espace vectoriel des matrices-colonnes réelles ou complexes à n éléments. Dans cet espace, les normes les plus courantes pour une matrice-colonne à éléments ξ^i, $i = 1, \ldots, n$, sont :

1) $\|\xi\|_1 = \sum_i |\xi^i|$, *norme octaédrique* ou *l-norme* ;

2) $\|\xi\|_2 = \left(\sum_i |\xi^i|^2\right)^{1/2}$, *norme euclidienne*, ou *norme unitaire* sur l'espace complexe ;

3) $\|\xi\|_p = \left(\sum_i |\xi^i|^p\right)^{1/p}$, $p > 1$, *norme de Hölder* ou *l_p-norme* ;

4) $\|\xi\|_\infty = \max_i |\xi^i|$, *norme cubique* ou *c-norme*.

Si l'on introduit dans l'espace arithmétique un produit scalaire en exigeant que la base formée des colonnes de la matrice unité soit orthonormée, on a pour tout x

$$\sqrt{(\xi, \xi)} = \left(\sum_i |\xi^i|^2\right)^{1/2}.$$

Il s'ensuit en vertu des propriétés du produit scalaire, que la norme euclidienne (resp. unitaire) vérifie tous les axiomes de la norme.

Le lecteur peut vérifier facilement les axiomes pour la l-norme et la c-norme. Pour la norme de Hölder, nous ne vérifierons pas les axiomes car cette norme ne se présentera pas dans notre exposé.

Dans un espace vectoriel quelconque, on peut utiliser les mêmes normes si l'on y fixe une base et qu'on associe à chaque vecteur l'une des normes mentionnées plus haut de sa colonne de coordonnées. Il va de soi qu'une norme ainsi construite dépend du choix de la base.

Pour toute norme sur l'espace vectoriel $\mathscr{L}_n$ on peut considérer un ensemble de transformations linéaires conservant cette norme, c'est-à-dire satisfaisant à la condition $\|A(x)\| = \|x\|$. Pour la norme euclidienne (resp. unitaire) ce sont des transformations orthogonales (resp. unitaires).

Appelons *sphère unité* dans un espace normé l'ensemble des vecteurs dont la norme est égale à l'unité.

PROPOSITION 1. *Etant donné la sphère unité dans un espace normé, on peut calculer la norme de tout vecteur.*

En effet, tout sous-espace unidimensionnel $\mathscr{L}_1$ coupe la sphère unité $\mathscr{S}$, car si $x \in \mathscr{L}_1$ on a $x_0 = \|x\|^{-1} x \in \mathscr{L}_1 \cap \mathscr{S}$. Par ailleurs, tout vecteur $y \in \in \mathscr{L}_1$ diffère du vecteur $x_0 \in \mathscr{S}$ par un facteur numérique. On peut maintenant noter que $\|y\| = \|\lambda x_0\| = |\lambda|$.

On a représenté fig. 52 les sphères unités dans l'espace bidimensionnel, qui correspondent à diverses normes. Les normes octaédrique et cubique s'appellent ainsi parce que pour $n = 3$ les sphères unités correspondantes sont respectivement un octaèdre et un cube. Laissons au lecteur le soin de le vérifier.

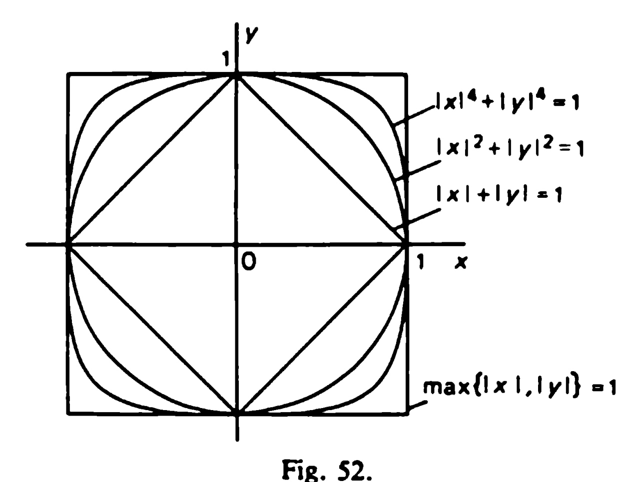


Fig. 52.

On comprend facilement que par exemple la norme octaédrique n'est engendrée par aucun produit scalaire. Sa sphère unité possède des propriétés de symétrie absolument différentes de celles d'une sphère ordinaire. Aussi la présence d'un produit scalaire engendrant cette norme contredit-elle le théorème de l'isomorphisme des espaces euclidiens. On n'entrera pas dans le détail de cette démonstration.

Considérons pour deux vecteurs x_1 et x_2 de l'espace normé un ensemble de tous les vecteurs de la forme

$$y = \alpha x_1 + (1 - \alpha) x_2,$$

où α est un nombre quelconque du segment [0, 1]. Cet ensemble de vecteurs s'appelle *segment d'extrémités* x_1 *et* x_2.

Quelle que soit la norme, la boule unité, c'est-à-dire l'ensemble des vecteurs dont la norme ne dépasse pas l'unité, possède la propriété suivante.

PROPOSITION 2. *Si les extrémités x_1 et x_2 d'un segment appartiennent à la boule unité, il en est de même de tout vecteur du segment.*

En effet, en vertu de l'axiome 3 de la norme, on a pour tout vecteur du segment

$$\|y\| \leqslant \alpha\|x_1\| + (1 - \alpha)\|x_2\|,$$

ce qui démontre l'assertion.

Les ensembles possédant la propriété énoncée par la proposition 2 sont appelés *convexes*. D'où la dénomination du troisième axiome.

Il ressort par exemple de la proposition 2 que l'astroïde, courbe d'équation $x^{2/3} + y^{2/3} = 1$, n'est le cercle unité pour aucune norme sur le plan.

3. Equivalence de normes. Soit φ une norme sur l'espace vectoriel $\mathscr{L}_n$. Si α est un nombre réel strictement positif, la fonction $\psi(x)$ telle que $\psi(x) = \alpha\varphi(x)$ pour tout $x \in \mathscr{L}_n$, est aussi une norme. Cette norme est notée $\alpha\varphi$.

Considérons deux normes φ et ψ. On dira que la norme φ *majore* la norme ψ et on écrira $\psi \leqslant \varphi$ si pour tout vecteur x de $\mathscr{L}_n$ est vérifiée l'inégalité $\psi(x) \leqslant \varphi(x)$. Il est évident que deux normes quelconques peuvent ne pas être liées par cette relation. Notons $O_\varphi(a, \varepsilon)$ le ε-voisinage du vecteur a pour la norme φ. Si $\psi \leqslant \varphi$, on vérifie aisément que $O_\varphi(a, \varepsilon) \subseteq O_\psi(a, \varepsilon)$ pour tous $a \in \mathscr{L}_n$ et $\varepsilon > 0$.

Les normes φ et ψ sont dites *équivalentes* s'il existe des nombres strictement positifs α_1 et α_2 tels que

$$\alpha_1 \varphi \leqslant \psi \leqslant \alpha_2 \varphi. \tag{2}$$

Il est évident que la relation d'équivalence est réfléxive, c'est-à-dire que chaque norme est équivalente à elle-même. Cette relation est également transitive : si $\alpha_1\varphi \leqslant \psi \leqslant \alpha_2\varphi$ et $\beta_1\psi \leqslant \chi \leqslant \beta_2\psi$, on a $\alpha_1\beta_1\varphi \leqslant \chi \leqslant$ $\leqslant \alpha_2\beta_2\varphi$. En outre, il découle de (2) que $\alpha_2^{-1}\psi \leqslant \varphi \leqslant \alpha_1^{-1}\psi$. Cela signifie que la relation d'équivalence des normes est symétrique.

Soient φ et ψ deux normes équivalentes. Il est aisé de démontrer que dans ce cas pour tous $a \in \mathscr{L}_n$ et $\varepsilon > 0$

$$O_\psi\left(a, \frac{\varepsilon}{\alpha_2}\right) \subseteq O_\varphi(a, \varepsilon) \subseteq O_\psi\left(a, \frac{\varepsilon}{\alpha_1}\right), \tag{3}$$

où les nombres α_1 et α_2 sont définis par la relation (2).

PROPOSITION 3. *La suite de vecteurs $\{x_k\}$ dans $\mathscr{L}_n$ converge en norme ψ si et seulement si elle converge en toute norme φ qui lui est équivalente.*

Démonstration. Posons que la suite $\{x_k\}$ converge vers le vecteur a en norme ψ. Démontrons qu'elle converge vers a en une norme équivalente φ. Choisissons un nombre arbitraire $\varepsilon > 0$. En vertu de la convergence en norme ψ, il existe un numéro $k_0(\varepsilon/\alpha_2)$ tel que $x_k \in O_\psi(a, \varepsilon/\alpha_2)$ pour tout $k \geqslant k_0(\varepsilon/\alpha_2)$. Il s'ensuit en vertu de (3) que $x_k \in O_\varphi(a, \varepsilon)$. Donc, le nombre $k_0(\varepsilon/\alpha_2)$ satisfait à la condition de convergence en norme φ.

L'assertion réciproque découle de la proposition démontrée en vertu de la symétrie de la relation d'équivalence entre les normes.

On démontre que dans un espace vectoriel de dimension finie toutes les normes sont équivalentes. La démonstration utilise les propriétés des fonctions continues (voir Koudriavtsev [21], t. I, § 19). Cependant, l'équivalence des trois normes principales peut être montrée sans difficulté.

Proposition 4. *Dans un espace arithmétique, la l-norme, la c-norme et la norme euclidienne sont équivalentes.*

Démonstration. Décomposons une matrice-colonne arbitraire $\boldsymbol{\xi}$ suivant les colonnes de la matrice unité. En appliquant les propriétés de la norme à cette décomposition, on obtient pour une norme arbitraire $\| * \|$

$$\|\boldsymbol{\xi}\| \leqslant \sum_i |\xi^i| \|\boldsymbol{e}_i\|. \tag{4}$$

Majorons le second membre en remplaçant chaque nombre $\|\boldsymbol{e}_i\|$ par sa valeur maximale. En désignant max $\|\boldsymbol{e}_i\|$ par α, on obtient

$$\|\boldsymbol{\xi}\| \leqslant \alpha \sum_i |\xi^i| = \alpha \|\boldsymbol{\xi}\|_1 . \tag{5}$$

Mais on peut aussi majorer le second membre de (4) en remplaçant chaque nombre $|\xi^i|$ par sa valeur maximale. On a

$$\|\boldsymbol{\xi}\| \leqslant \max_i |\xi^i| \sum_i \|\boldsymbol{e}_i\| = \beta \|\boldsymbol{\xi}\|_\infty , \tag{6}$$

où $\beta = \sum \|\boldsymbol{e}_i\|$. Appliquons (5) à la c-norme, et (6) à la l-norme. Alors, puisque $\alpha \neq 0$, on a

$$\alpha^{-1} \|\boldsymbol{\xi}\|_\infty \leqslant \|\boldsymbol{\xi}\|_1 \leqslant \beta \|\boldsymbol{\xi}\|_\infty ,$$

ce qui démontre l'équivalence de la l-norme et de la c-norme.

L'estimation

$$\max_i |\xi^i| \leqslant \left(\sum_i |\xi^i|^2 \right)^{1/2} \leqslant \sqrt{n} \max_i |\xi^i|$$

entraîne l'équivalence de la c-norme et de la norme euclidienne. La proposition est démontrée. On a démontré en même temps la proposition suivante.

PROPOSITION 5. *Toute norme sur l'espace arithmétique est majorée par le produit de la l-norme par le nombre α,*
le produit de la c-norme par le nombre β,
le produit de la norme euclidienne par le nombre β.

PROPOSITION 6. *Toute norme φ est une fonction continue en chacune des trois normes fondamentales, c'est-à-dire que $\varphi(\xi_k) \to \varphi(\xi_0)$ pour toute suite $\{\xi_k\}$ convergeant vers ξ_0 en l-norme, c-norme ou en norme euclidienne.*

Il suffit de le démontrer par exemple pour la c-norme. Selon (1) on a

$$|\varphi(\xi_k) - \varphi(\xi_0)| \leqslant \varphi(\xi_k - \xi_0) \leqslant \beta \|\xi_k - \xi_0\|_\infty .$$

Vu que $\|\xi_k - \xi_0\|_\infty \to 0$, la proposition 6 découle des propriétés des suites numériques.

PROPOSITION 7. *Pour toute norme φ il existe des nombres strictement positifs ρ_1 et ρ_2 tels que chaque matrice-colonne ξ pour laquelle $\|\xi\|_2 = 1$ vérifie la double inégalité $\rho_1 \leqslant \varphi(\xi) \leqslant \rho_2$.*

Pour le démontrer, profitons des résultats fournis par l'analyse mathématique (voir Koudriavtsev [21], t. I, § 19). La sphère unité euclidienne, ensemble de matrices-colonnes telles que $\|\xi\|_2 = 1$, est fermée et bornée en norme euclidienne. La fonction φ est continue en norme euclidienne et, par suite, atteint sur la sphère les valeurs maximale et minimale. En notant ρ_2 et ρ_1 ces valeurs, on aboutit à la double inégalité cherchée. Il reste à démontrer que $\rho_1 > 0$. Or c'est évident, car il existe une matrice-colonne ξ_0 pour laquelle $\varphi(\xi_0) = \rho_1$ et $\|\xi_0\|_2 = 1$.

On est maintenant en mesure de démontrer le théorème suivant.

THÉORÈME 1. *Dans un espace arithmétique, toutes les normes sont équivalents.*

DÉMONSTRATION. On démontrera l'équivalence de toute norme φ à la norme euclidienne. D'où, en vertu de la transitivité de la relation d'équivalence des normes, découlera l'assertion du théorème.

Faisons correspondre à une matrice-colonne arbitraire $\xi \neq \mathbf{0}$ une matrice-colonne ξ^0 telle que $\xi = \|\xi\|_2 \xi^0$. Alors, $\varphi(\xi) = \|\xi\|_2 \varphi(\xi^0)$ et selon la proposition 7 on peut écrire

$$\rho_1 \|\xi\|_2 \leqslant \varphi(\xi) \leqslant \rho_2 \|\xi\|_2 .$$

Vu que pour $\xi = \mathbf{0}$ l'inégalité est évidente, le théorème est démontré.

Etant donné que toutes les normes sont équivalentes, on reste parfois indifférent au choix de la norme, si bien que le lecteur pourra rencontrer des énoncés qui contiennent par exemple une expression : « si l'inégalité donnée est vérifiée en une norme quelconque... ». Or dans les calculs, il est

souvent fort important à quelle norme on a affaire, et la possibilité de choisir librement la norme qui convient le plus, simplifie fortement les raisonnements.

4. Normes des matrices. Considérons l'espace vectoriel $\mathscr{M}_{m,n}$ des matrices à m lignes et n colonnes. Comme dans tout espace vectoriel, on peut y introduire différentes normes. On s'intéressera à celles d'entre elles qui sont rattachées aux normes des matrices-colonnes, ainsi qu'au fait que les éléments de l'espace sont des matrices.

L'exposé qui suit traite des matrices réelles. Pour les matrices complexes, les différences qui surgissent sont évidentes et peu importantes. Par exemple, on remplace la norme euclidienne par la norme unitaire, les matrices orthogonales, par les matrices unitaires.

DÉFINITION. On dit que la norme sur $\mathscr{M}_{m,n}$ est *compatible* ou *concordante* avec les normes sur les espaces arithmétiques $\mathscr{R}_m$ et $\mathscr{R}_n$ si pour toute matrice A et toute matrice-colonne $\xi \in \mathscr{R}_n$ on a

$$\|A\xi\| \leqslant \|A\|\|\xi\|. \tag{7}$$

$A\xi$ et ξ sont ici des matrices-colonnes à m et n éléments respectivement, dont les normes sont choisies dans les espaces $\mathscr{R}_m$ et $\mathscr{R}_n$.

Montrons sur un exemple qu'il existe des normes compatibles. A cet effet, considérons la fonction d'une matrice $A \in \mathscr{M}_{m,n}$:

$$\varphi(A) = \sup \frac{\|A\xi\|}{\|\xi\|}. \tag{8}$$

Vu que $\|A\xi\|/\|\xi\| = \|A\xi^0\|$, où $\xi^0 = \|\xi\|^{-1}\xi$, la fonction φ peut être écrite également sous la forme

$$\varphi(A) = \sup_{\|\xi\|=1} \|A\xi\|. \tag{9}$$

PROPOSITION 8. *La fonction* (8) *est une norme sur* $\mathscr{M}_{m,n}$, *qui est compatible avec toutes les normes sur* $\mathscr{R}_m$ *et* $\mathscr{R}_n$.

L'existence de la borne supérieure sera établie si l'on démontre que le rapport $\|A\xi\|/\|\xi\|$ est majoré. En utilisant le théorème 1, on peut écrire

$$\|A\xi\| \leqslant \alpha\|A\xi\|_1 = \alpha \sum_i \sum_j a^i_j\xi^j \leqslant \alpha \max_j |\xi^j| \sum_{ij} |a^i_j| = \\ = \beta\|\xi\|_\infty \leqslant \gamma\|\xi\|.$$

D'où $\|A\xi\|/\|\xi\| \leqslant \gamma$.

Vérifions les axiomes de la norme.

1) L'expression $\|A\xi\|/\|\xi\|$ est positive et, par suite $\varphi(A) \geqslant 0$. Ceci étant, $\varphi(A) = 0$ si et seulement si $\|A\xi\| = 0$ pour tous les ξ. Or la condi-

tion $\|A\xi\| = 0$ est équivalente à $A\xi = \mathbf{0}$ et est satisfaite pour tous les ξ si et seulement si $A = O$.

2) L'identité $(\lambda A)\xi = \lambda(A\xi)$ entraîne $\|(\lambda A)\xi\| = |\lambda| \cdot \|A\xi\|$. Il n'est pas difficile de démontrer, en utilisant la définition de la borne supérieure, que la multiplication de tous les éléments de l'ensemble par un nombre positif entraîne celle de la borne supérieure de l'ensemble par ce nombre : $\sup_{\|\xi\|=1} |\lambda| \|A\xi\| = |\lambda| \sup_{\|\xi\|=1} \|A\xi\|$. On a ainsi démontré que la fonction $\varphi(A)$ possède la propriété d'homogénéité positive.

3) Pour tous A, B et $\|\xi\| = 1$, on a

$$\|(A + B)\xi\| = \|A\xi + B\xi\| \leqslant \|A\xi\| + \|B\xi\|.$$

On peut aussi démontrer la propriété générale de la borne supérieure, soit :

$$\sup_{p \in P,\, q \in Q} (p + q) = \sup_P (p) + \sup_Q (q),$$

si P et Q sont des ensembles de nombres réels (voir Koudriavtsev [21], t. I, p. 26, ex. 1).

Ainsi donc, la fonction (8) est une norme. Démontrons qu'elle est compatible avec les autres normes. En effet, pour $\xi \neq \mathbf{0}$ on a par définition de la borne supérieure $\|A\xi\|/\|\xi\| \leqslant \varphi(A)$, d'où $\|A\xi\| \leqslant \varphi(A)\|\xi\|$.

Définition. La norme sur $\mathscr{M}_{m,n}$ définie par la formule (8) s'appelle *norme induite par les normes sur les espaces* $\mathscr{R}_m$ et $\mathscr{R}_n$, ou tout simplement, *norme induite*.

Proposition 9. *Toute norme compatible majore la norme induite.*

En effet, toute norme compatible de la matrice A majore le rapport $\|(A\xi\|/\|\xi\|$. Or la borne supérieure est le plus petit des majorants, de sorte que la norme induite de la matrice A ne dépasse aucune de ses normes compatibles.

Définissons deux propriétés importantes des normes de matrices. Si une norme sur l'espace des matrices carrées $\mathscr{M}_{n^2}$ est telle que $\|E\| = 1$, on dit qu'elle « conserve l'unité ». Il est aisé de voir que pour toute norme φ sur $\mathscr{M}_{n^2}$ il existe une norme de la forme $c\varphi$ et une seule qui possède cette propriété.

Les normes sur l'espace $\mathscr{M}_{n^2}$ qui satisfont pour tous A et B à la condition

$$\|AB\| \leqslant \|A\| \cdot \|B\|, \tag{10}$$

sont dites *matricielles* ou *annulaires*. On utilisera le dernier terme, quoique moins employé, pour éviter des expressions ambiguës quand une norme de la matrice n'est pas une norme matricielle. La condition (10) sera appelée *propriété annulaire* d'une norme.

On constate facilement que pour des normes annulaires on a $\|A\| \leqslant \leqslant \|E\| \cdot \|A\|$, d'où

$$\|E\| \geqslant 1. \tag{11}$$

En outre, il découle de la propriété annulaire que

$$\|A^k\| \leqslant \|A\|^k \tag{12}$$

pour tout k entier naturel, et que

$$\|A^{-1}\| \geqslant \|A\|^{-1}. \tag{13}$$

Démontrons maintenant la proposition suivante.

PROPOSITION 10. *Toute norme induite sur $\mathcal{M}_{n^2}$ conserve l'unité et possède la propriété annulaire.*

La première partie de la proposition est évidente. La seconde découle de la majoration suivante où on utilise encore une fois la propriété de la borne supérieure, appliquée lors de la démonstration de la proposition 8 :

$$\|AB\| = \sup_{\|\xi\|=1} \|A(B\xi)\| \leqslant \sup_{\|\xi\|=1} \|A\| \cdot \|B\xi\| = \|A\| \sup_{\|\xi\|=1} \|B\xi\| =$$
$$= \|A\| \cdot \|B\|.$$

Fixons l'attention sur le fait que l'inégalité (10) peut aussi se vérifier dans un cas plus général. Il suffit d'admettre que les matrices A et B soient rectangulaires et que leur produit soit défini (par exemple, $A \in \mathcal{M}_{m,n}$, $B \in \in \mathcal{M}_{n,l}$). Dans ce cas, les normes sur trois espaces matriciels $\mathcal{M}_{m,n}$, $\mathcal{M}_{n,l}$ et $\mathcal{M}_{m,l}$ peuvent être rendues compatibles de façon à avoir

$$\|AB\|_{\mathrm{III}} \leqslant \|A\|_{\mathrm{I}} \cdot \|B\|_{\mathrm{II}}. \tag{14}$$

Cette propriété de compatibilité sera également appelée *propriété annulaire*. En particulier, pour $l = 1$, la propriété (14) marque la compatibilité de la norme $\| * \|_{\mathrm{I}}$ avec les normes $\| * \|_{\mathrm{II}}$ et $\| * \|_{\mathrm{III}}$ sur les espaces arithmétiques.

En examinant la démonstration de la proposition 10, on s'aperçoit qu'on peut énoncer la proposition suivante.

PROPOSITION 11. *L'inégalité* (14) *se vérifie pour toutes matrices de dimensions appropriées si les normes $\| * \|_{\mathrm{III}}$ et $\| * \|_{\mathrm{II}}$ sont induites, et que la norme $\| * \|_{\mathrm{I}}$ soit compatible.*

En rapport avec la définition de la norme induite, un intérêt naturel apparaît pour la borne inférieure

$$g(A) = \inf_{\xi \neq 0} \frac{\|A\xi\|}{\|\xi\|}. \tag{15}$$

Elle existe car l'ensemble des nombres de la forme $\|A\xi\|/\|\xi\|$ est minoré par zéro.

Supposons que A est une matrice carrée et qu'on prend pour les matrices-colonnes $\|\xi\|$ et $\|A\xi\|$ la même norme. On a alors la

PROPOSITION 12. *Soit* $\| * \|$ *la norme induite. Si* $\det A \neq 0$, $g(A) = (\|A^{-1}\|)^{-1}$. *Mais si* $\det A = 0$, $g(A) = 0$.

DÉMONSTRATION. Si $\det A \neq 0$, l'expression (15) peut être mise sous la forme $\inf (\|\eta\|/\|A^{-1}\eta\|)$, où $\eta = A\xi$. Remarquons que $\eta \neq \mathbf{0}$. Il faut ensuite se référer à la propriété suivante des bornes inférieures et supérieures d'ensembles numériques : si P est un ensemble de nombres strictement positifs et Q l'ensemble de tous les nombres de la forme p^{-1}, où $p \in P$, on a $\inf Q = (\sup P)^{-1}$. Laissons au lecteur le soin de le vérifier. Il s'ensuit

$$g(A) = \left(\sup_{\eta \neq 0} \frac{\|A^{-1}\eta\|}{\|\eta\|} \right)^{-1},$$

de sorte que la première assertion est démontrée.

La seconde assertion est évidente car $\inf \|A\xi\|/\|\xi\| \geqslant 0$. Mais pour $\det A = 0$, il existe une matrice-colonne non nulle ξ_0 pour laquelle $A\xi_0 = \mathbf{0}$.

5. Normes de matrices les plus usuelles. Voyons quelles normes sont induites sur l'espace des matrices par les normes sur les espaces arithmétiques que nous avons passées en revue au point 1.

a) Supposons que dans les espaces arithmétiques on a choisi les normes euclidiennes (ou unitaires). La norme induite par ces dernières est appelée *norme spectrale* de la matrice. C'est

$$\sup_{\xi \neq 0} \frac{\|A\xi\|}{\|\xi\|}.$$

Dans la proposition 10 du § 2 on a vu que pour les matrices carrées cette borne supérieure est atteinte et qu'elle est égale au nombre singulier maximal de la matrice A. On verra plus bas que cette propriété est aussi vérifiée pour les matrices rectangulaires, mais d'abord démontrons la proposition suivante.

PROPOSITION 13. *La norme spectrale d'une matrice ne varie pas quand on multiplie cette matrice à droite ou à gauche par une matrice orthogonale* (*resp. unitaire*).

DÉMONSTRATION. La multiplication d'une matrice-colonne par une matrice orthogonale ne modifie pas sa norme euclidienne : $\|U\xi\| = \|\xi\|$. Il en découle que la norme spectrale d'une matrice orthogonale (resp. unitaire) vaut 1. Maintenant, en vertu de la proposition 4, on peut écrire

$$\|UA\| \leqslant \|U\| \cdot \|A\| = \|A\|$$

et
$$\|A\| = \|U^{-1}UA\| \leqslant \|U^{-1}\| \cdot \|UA\| = \|UA\|,$$
ce qui démontre l'assertion pour les facteurs à gauche. Pour les facteurs à droite la démonstration est presque la même.

Pour calculer une norme spectrale, on peut se servir de la proposition 13 et du théorème 1m, § 1, qui permettent de réduire le calcul de la norme spectrale d'une matrice A de type (m, n) à celui d'une matrice A' de la forme (comp. (17), § 1) :
$$A' = \begin{Vmatrix} D_r & O \\ O & O \end{Vmatrix}.$$
D_r est ici une matrice carrée diagonale d'ordre $r = \mathrm{Rg}\, A$, sur la diagonale de laquelle se trouvent les nombres singuliers non nuls α_i de la matrice A. Ceci étant, ils sont numérotés de façon à avoir $\alpha_1 \geqslant \alpha_2 \geqslant \ldots \geqslant \alpha_r$.

Considérons une matrice-colonne $\boldsymbol{\xi} \in \mathscr{R}_n$. Pour cette dernière,
$$\|\boldsymbol{\xi}\| = \left(\sum_{i=1}^{n} (\xi^i)^2\right)^{1/2} \quad \text{et} \quad \|A'\boldsymbol{\xi}\| = \left(\sum_{i=1}^{r} (\alpha_i \xi^i)^2\right)^{1/2}.$$
D'où, en remplaçant $\alpha_2, \ldots, \alpha_r$ par α_1, on obtient $\|A'\boldsymbol{\xi}\|/\|\boldsymbol{\xi}\| \leqslant \alpha_1$, l'égalité étant vérifiée si $\boldsymbol{\xi} = {}^t\|1, 0, \ldots, 0\|$. Ainsi donc, on a démontré la

PROPOSITION 14. *La norme spectrale d'une matrice est égale à son nombre singulier maximal* : $\|A\| = \alpha_1$.

Pour une matrice carrée régulière, on obtient de la proposition 12 et de la proposition 10 du § 2, que
$$\|A^{-1}\| = \alpha_n^{-1}, \tag{16}$$
où α_n est le nombre singulier minimal (pour $\det A \neq 0$, il n'est pas nul).

b) Admettons qu'on a choisi les l-normes sur $\mathscr{R}_m$ et $\mathscr{R}_n$. Majorons $\|A\boldsymbol{\xi}\|_1$ pour la matrice-colonne $\boldsymbol{\xi} \neq \mathbf{0}$:
$$\|A\boldsymbol{\xi}\|_1 = \sum_j \left| \sum_k a^j_k \xi^k \right| \leqslant \sum_k \left(\sum_j |a^j_k|\right) |\xi^k| \leqslant$$
$$\leqslant \left(\max_k \sum_j |a^j_k|\right) \sum_j |\xi^j|.$$
Ainsi donc,
$$\frac{\|A\boldsymbol{\xi}\|_1}{\|\boldsymbol{\xi}\|_1} \leqslant \max_k \sum_j |a^j_k|.$$
Montrons que l'égalité est possible ici. Supposons que le maximum est atteint pour la valeur $k = s$. Choisissons pour matrice-colonne $\boldsymbol{\xi}$ la s-ième

colonne de la matrice unité d'ordre n. $A\boldsymbol{\xi}$ est alors la s-ième colonne de la matrice A et sa l-norme vaut

$$\|Ae_s\|_1 = \sum_j |a_s^j|.$$

Vu que $\|e_s\|_1 = 1$, on a $\|Ae_s\|_1 / \|e_s\|_1 = \sum_j |a_s^j| = \max_k \sum_j |a_k^j|$. Donc, la norme induite par les l-normes sur les espaces arithmétiques est la norme

$$\|A\|_l = \max_k \sum_j |a_k^j|,$$

c'est-à-dire la somme maximale parmi les sommes des modules des éléments de chaque colonnes de la matrice, autrement dit le maximum des l-normes des matrices-colonnes.

c) Supposons qu'on a choisi les c-normes sur $\mathscr{R}_m$ et $\mathscr{R}_n$. Majorons $\|A\boldsymbol{\xi}\|_\infty$ pour la matrice-colonne $\boldsymbol{\xi} \neq \mathbf{0}$:

$$\|A\boldsymbol{\xi}\|_\infty = \max_j \left| \sum_k a_k^j \xi^k \right| \leqslant \max_j \sum_k |a_k^j| |\xi^k| \leqslant$$
$$\leqslant \left(\max_i |\xi^i|\right) \max_j \sum_k |a_k^j|.$$

D'où

$$\frac{\|A\boldsymbol{\xi}\|_\infty}{\|\boldsymbol{\xi}\|_\infty} \leqslant \max_j \sum_k |a_k^j|.$$

Montrons que l'égalité est ici possible. Posons que le maximum est atteint pour la valeur $k = s$. On exige que le s-ième élément de la matrice-colonne $\boldsymbol{\eta} = A\boldsymbol{\xi}$ soit égal à

$$\sum_k |a_k^s|.$$

Il suffit pour cela de choisir $\boldsymbol{\xi}$ de manière que pour tous les k soit vérifiée l'égalité $a_k^s \xi^k = |a_k^s|$. Il est évident que c'est toujours réalisable et que de plus $\|\boldsymbol{\xi}\|_\infty = 1$. Ensuite, pour toute matrice-colonne on a $\max_j |\eta^j| \geqslant \eta^s$, de sorte que

$$\max_j |\eta^j| \geqslant \sum_k |a_k^s| = \max_j \sum_k |a_k^j|,$$

ce qui avec la majoration précédente donne

$$\sup_{\xi} \frac{\|A\xi\|_\infty}{\|\xi\|_\infty} = \max_j \sum_k |a_k^j|.$$

Donc, la norme induite par les c-normes sur les espaces arithmétiques est la norme

$$\|A\|_c = \max_j \sum_k |a_k^j|,$$

c'est-à-dire la somme maximale parmi les sommes des modules des éléments de chaque ligne de la matrice.

On peut mettre toute matrice à m lignes et n colonnes sous forme de matrice-colonne à mn éléments et considérer pour ces matrices les mêmes normes qu'on a définies pour les matrices-colonnes au point 2. Dans ce cas, n'étant pas induites, ces normes ne possèdent pas obligatoirement les propriétés utiles des normes induites.

d) Norme *euclidienne* de la matrice réelle et norme *unitaire* de la matrice complexe. On pose par définition

$$\|A\|_E = \left(\sum_{i,j} |a_j^i|^2\right)^{1/2}.$$

L'élément de la matrice ${}^tAA$ rangé dans la i-ième ligne et la j-ième colonne est

$$\sum_k a_{ki} a_{kj}.$$

On en tire que le carré de $\|A\|_E$ est égale à la trace de ${}^tAA$, c'est-à-dire que

$$\|A\|_E = \sqrt{\operatorname{tr} {}^tAA}.$$

De façon analogue, on démontre que la norme unitaire d'une matrice complexe est

$$\|A\|_U = \sqrt{\operatorname{tr} {}^t\bar{A}A}.$$

Dans la suite, on ne traitera que des matrices réelles en laissant au lecteur le soin de reporter les résultats obtenus sur les matrices complexes.

L'expression obtenue pour $\|A\|_E$ justifie la proposition suivante.

PROPOSITION 15. *La norme euclidienne de la matrice est égale à la racine carrée de la somme des carrés de ses nombres singuliers.*

PROPOSITION 16. *La norme euclidienne de la matrice A ne varie pas quand on multiplie A à droite ou à gauche par une matrice orthogonale.*

Démonstration. Si U est une matrice orthogonale, on a

$$\operatorname{tr} {}^t(UA)UA = \operatorname{tr} {}^tA{}^tUUA = \operatorname{tr} {}^tAA.$$

La norme euclidienne étant invariante par transposition, on a aussi

$$\|AU\|_E = \|{}^t(AU)\|_E = \|{}^tU{}^tA\|_E = \|{}^tA\|_E = \|A\|_E.$$

On achève ainsi la démonstration de la proposition.

La norme euclidienne possède la propriété annulaire : si le produit des matrices A et B est défini, $\|AB\|_E \leqslant \|A\|_E\|B\|_E$.

En effet, en vertu de l'inégalité de Cauchy-Bouniakovski,

$$\left(\sum_j a_{ij}b_{jk}\right)^2 \leqslant \sum_j a_{ij}^2 \sum_j b_{jk}^2$$

pour tous i et k. En sommant ces inégalités suivant tous i et k, on obtient $\|AB\|_E^2 \leqslant \|A\|_E^2\|B\|_E^2$, d'où l'expression nécessaire.

Il s'ensuit, en particulier, que la norme euclidienne est compatible avec les normes euclidiennes sur les espaces arithmétiques :

$$\|A\xi\|_2 \leqslant \|A\|_E\|\xi\|_2.$$

Il n'existe pas de norme sur l'espace arithmétique qui puisse induire une norme matricielle euclidienne. En effet, $\|E\|_E = \sqrt{n}$.

e) La norme $\|A\| = \max_{i,j} |a_j^i|$ ne possède pas la propriété annulaire tout en conservant l'unité. On utilise beaucoup plus souvent dans l'espace des matrices carrées d'ordre n la norme suivante

$$\|A\|_{c^\cdot} = n \max_{i,j} |a_j^i|.$$

Elle possède la propriété annulaire et est compatible avec les trois normes principales sur les espaces arithmétiques. Démontrons-le.

L'élément du produit AB peut être majoré en module de la façon suivante :

$$\left|\sum_j a_j^i b_k^j\right| \leqslant \sum_j |a_j^i||b_k^j| \leqslant \max_{i,j} |a_j^i| \sum_j |b_k^j| \leqslant$$

$$\leqslant n \max_{i,j} |a_j^i| \max_{j,k} |b_k^j|.$$

Cette majoration a lieu également pour l'élément de AB maximal en module. En multipliant les deux membres de l'inégalité par n, on obtient $\|AB\|_{c^\cdot} \leqslant \|A\|_{c^\cdot} \cdot \|B\|_{c^\cdot}$. On a démontré ainsi la propriété annulaire.

Vu que $n^2 (\max_{i,j} |a_{ij}|)^2$ est supérieur à la somme des carrés de tous les éléments de la matrice, la norme considérée est supérieure à la norme eucli-

dienne. On a vu que la norme matricielle euclidienne est compatible avec la norme euclidienne sur $\mathscr{R}_n$. Par suite, la norme $\| * \|_{c'}$ est aussi compatible avec la norme euclidienne sur $\mathscr{R}_n$.

D'une façon analogue, la norme $\| * \|_{c'}$ est supérieure aux normes induites par la l-norme et la c-norme, et, par suite, est compatible avec la l-norme et la c-norme sur $\mathscr{R}_n$.

f) La norme $\sum_{i,j} |a_j^i|$ ne nous intéressera pas. Notons que la norme $\frac{1}{n} \sum_{i,j} |a_j^i|$ conserve l'unité.

6. Convergence en éléments. On étudie pour les suites dans l'espace arithmétique la convergence en éléments (ou en coordonnées). La suite $\{\xi_k\}$ *converge en éléments* vers ξ_0 si les suites formées d'éléments des matrices-colonnes ξ_k convergent vers les éléments correspondants de la matrice-colonne ξ_0 :

$$\lim_{k \to \infty} \xi_k^i = \xi_0^i, \quad i = 1, \ldots, n.$$

D'une façon plus détaillée, cela signifie que quels que soient les nombres strictement positifs $\varepsilon_1, \ldots, \varepsilon_n$ il existe des numéros $k_1(\varepsilon_1), \ldots, k_n(\varepsilon_n)$ tels que pour tous les $i = 1, \ldots, n$ l'inégalité $|\xi_k^i - \xi_0^i| < \varepsilon_i$ se vérifie si $k > k_i(\varepsilon_i)$.

On remarque aussitôt que la convergence en éléments se confond avec la convergence en c-norme. En effet, si la suite converge en éléments, choisissons un $\varepsilon > 0$ quelconque et posons $\varepsilon_1 = \ldots = \varepsilon_n = \varepsilon$. Dans ce cas, on aura alors pour $k > \max_i k_i(\varepsilon)$ l'inégalité $\max |\xi_k^i - \xi_0^i| < \varepsilon$, ce qui équivaut à la convergence en c-norme. L'assertion réciproque est aussi évidente.

Du théorème 1 et de la proposition 3 il découle maintenant que la suite converge en éléments si et seulement si elle converge en une norme quelconque.

Tout ce qui a été dit plus haut se rapporte aussi à la convergence en éléments d'une suite de matrices. En écrivant les éléments de la matrice de type (m, n) en une colonne à mn éléments, on peut considérer l'espace $\mathscr{M}_{m,n}$ comme un espace arithmétique et formuler le résultat suivant.

La suite de matrices A_k converge vers la matrice A_0 en éléments si et seulement si elle converge vers A_0 en une norme quelconque.

CHAPITRE XII

THÉORÈME DE JORDAN. FONCTIONS DE MATRICES

§ 1. Polynômes annulateurs

1. Divisibilité des polynômes. On rappellera ici les propriétés élémentaires de divisibilité des polynômes d'une variable, nécessaires à l'exposé ultérieur. Les polynômes à coefficients complexes (resp. réels) constituent un espace vectoriel complexe (resp. réel) relativement aux opérations habituelles d'addition et de multiplication par un nombre. L'élément nul de cet espace est le polynôme identiquement, nul, autrement dit le polynôme dont tous les coefficients sont nuls. On l'appellera plus loin polynôme nul ou polynôme égal à zéro.

Outre les opérations linéaires, l'ensemble des polynômes est muni de l'opération de multiplication. Notons que la multiplication par un nombre se confond avec la multiplication par un polynôme de degré nul dont le terme constant est égal à ce nombre.

Soient les polynômes q et p. Admettons qu'il existe des polynômes h et r tels que $q = hp + r$ et le degré de r est strictement inférieur à celui de p. On dit alors que h est le *quotient* de q par p, et r est le *reste*.

Si le reste de la division de q par p vaut zéro, c'est-à-dire s'il existe un polynôme h tel que $q = hp$, on dit que q est *divisible* par p ou que p est le *diviseur* de q.

PROPOSITION 1 (THÉORÈME DE BÉZOUT). *Le polynôme p est divisible par le binôme $t - \mu$ si et seulement si $p(\mu) = 0$.*

Le reste r de la division de p par $t - \mu$ est de degré strictement inférieur à celui de $t - \mu$, c'est-à-dire est une constante. En portant μ dans l'égalité $p = h(t - \mu) + r$, on trouve $r = p(\mu)$, d'où la proposition à démontrer.

DÉFINITION. On dit que le polynôme d est *le plus grand commun diviseur* des polynômes $p_1, \ldots, p_s$ s'il divise chacun d'eux et se divise par tout autre diviseur commun à ces polynômes.

Le plus grand commun diviseur des polynômes $p_1, \ldots, p_s$ est noté PGCD $(p_1, \ldots, p_s)$. Les polynômes $p_1, \ldots, p_s$ sont dits *premiers entre eux* si le PGCD $(p_1, \ldots, p_s)$ est de degré 0.

PROPOSITION 2. *Pour que les polynômes complexes $p_1, \ldots, p_s$ ne soient*

pas premiers entre eux il faut et il suffit qu'ils possèdent une racine commune.

La nécessité de la condition est évidente, tandis que la suffisance découle directement du théorème de Bézout.

PROPOSITION 3. *Soient $p_1, \ldots, p_s$ des polynômes dont l'un au moins est différent de zéro et d = PGCD $(p_1, \ldots, p_s)$. Il existe alors des polynômes $u_1, \ldots, u_s$ tels que*

$$d = u_1 p_1 + \ldots + u_s p_s. \tag{1}$$

DÉMONSTRATION. Notons $\mathscr{I}$ l'ensemble de tous les polynômes de la forme $f_1 p_1 + \ldots + f_s p_s$, où $f_1, \ldots, f_s$ sont des polynômes. $\mathscr{I}$ possède les propriétés suivantes :

a) Si $g, h \in \mathscr{I}$, on a $g + h \in \mathscr{I}$.

b) Si $g \in \mathscr{I}$, alors $hg \in \mathscr{I}$ quel que soit le polynôme h.

Il va de soi que chacun des polynômes $p_1, \ldots, p_s$ appartient à $\mathscr{I}$. Il s'ensuit en particulier que $\mathscr{I}$ contient des polynômes différents de zéro.

Dans tout ensemble contenant au moins un polynôme non nul, il y a un polynôme non nul de degré minimal. En effet, les degrés des polynômes sont des nombres entiers positifs et par suite, dans l'ensemble de degrés des polynômes non nuls il existe un élément minimal. Soit $d = u_1 p_1 + \ldots + u_s p_s$ un polynôme non nul de $\mathscr{I}$ de degré minimal. Alors d est le diviseur de tout polynôme g de $\mathscr{I}$. En effet, soit $g = hd + r$ et $r \neq 0$. Dans ce cas, $r = g - hd$ et, par suite, r se trouve dans $\mathscr{I}$ avec g et d. Or le degré de r est strictement inférieur à celui de d, ce qui contredit le choix de d. Donc, $r = 0$ et d est le diviseur de tout polynôme de $\mathscr{I}$, en particulier des polynômes $p_1, \ldots, p_s$.

Supposons maintenant que d_1 est un diviseur commun aux polynômes $p_1, \ldots, p_s$. Alors d_1 est le diviseur de tous les termes du second membre de l'égalité (1) et donc le diviseur de d, ce qui achève la démonstration.

COROLLAIRE. *Les polynômes $p_1, \ldots, p_s$ sont premiers entre eux si et seulement s'il existe des polynômes $u_1, \ldots, u_s$ tels que*

$$u_1 p_1 + \ldots + u_s p_s = 1.$$

DÉFINITION. Soient donnés des polynômes non nuls $p_1, \ldots, p_s$. On dit que le polynôme q est un *multiple commun* à ces polynômes s'il est divisible par chacun d'eux. Le multiple commun aux polynômes $p_1, \ldots, p_s$ s'appelle *le plus petit commun multiple* s'il est le diviseur de tout autre multiple commun à ces polynômes.

Le plus petit commun multiple des polynômes $p_1, \ldots, p_s$ est noté PPCM $(p_1, \ldots, p_s)$.

PROPOSITION 4. *Le multiple commun q des polynômes $p_1, \ldots, p_s$ est leur*

plus petit commun multiple si et seulement si les quotients de q par $p_1, \ldots, p_s$ sont premiers entre eux.

DÉMONSTRATION. Pour tous les $i = 1, \ldots, s$, les quotients h_i sont définis par les égalités $q = h_i p_i$. Si $d = \text{PGCD}(h_1, \ldots, h_s)$ est un polynôme de degré non nul, il existe des polynômes $f_1, \ldots, f_s$ tels que pour tous les $i = 1, \ldots, s$ sont vérifiées les égalités $q = f_i d p_i$. Donc, q se divise par d. En désignant le quotient par q', on a $f_1 p_1 = \ldots = f_s p_s = q'$. Cela signifie que q' est un multiple commun aux polynômes $p_1, \ldots, p_s$. Son degré est strictement inférieur à celui de q et, par suite, il ne se divise pas par q. Donc, q n'est pas le plus petit commun multiple.

Inversement, soit q^* un multiple commun aux polynômes $p_1, \ldots, p_s$ mais non pas le plus petit. Son degré est alors strictement supérieur à celui de $q = \text{PPCM}(p_1, \ldots, p_s)$. En effet, q^* se divise par q et par suite, son degré ne peut être strictement inférieur à celui de q. Si les degrés de q^* et de q étaient égaux, alors pour la même raison q^* différerait de q par un facteur numérique. Dans ce cas, q^* serait de même le plus petit commun multiple contrairement à l'hypothèse.

Ainsi donc, il existe un polynôme f de degré non nul, tel que $q^* = fq$. On peut maintenant écrire pour chaque i que $q^* = fh_i p_i$, où h_i est le quotient de q par p_i. Il en découle que les quotients de q^* par p_i possèdent un diviseur commun de degré strictement positif. La proposition est démontrée.

2. Polynômes de transformations. Considérons un espace vectoriel $\mathcal{L}_n$ de dimension n sans préciser pour l'instant si c'est un espace réel ou complexe. Rappelons (voir point 6, § 3, ch. VI) que pour les transformations linéaires de l'espace $\mathcal{L}_n$ on a défini les opérations d'addition, de multiplication par un nombre et de multiplication des transformations. En particulier, on a défini l'opération d'élévation d'une transformation linéaire à une puissance positive entière. Introduisons une définition complémentaire selon laquelle chaque transformation linéaire de puissance nulle est égale à la transformation identique $\mathbf{E}$ de l'espace $\mathcal{L}_n$.

Cela nous permet de porter une transformation linéaire $\mathbf{A}$, en tant qu'une valeur de la variable t, dans tout polynôme p :

$$p(\mathbf{A}) = \alpha_0 \mathbf{E} + \alpha_1 \mathbf{A} + \alpha_2 \mathbf{A}^2 + \ldots + \alpha_m \mathbf{A}^m.$$

Il va de soi que la valeur obtenue $p(\mathbf{A})$ du polynôme p sera aussi une transformation linéaire de l'espace $\mathcal{L}_n$. L'image d'un vecteur x par cette transformation sera notée $p(\mathbf{A})(x)$ ou $p^{\mathbf{A}}(x)$. La valeur du polynôme sur la transformation $\mathbf{A}$ est appelée tout simplement *polynôme de la transformation* $\mathbf{A}$.

L'espace $\mathcal{L}_n$ étant rapporté à une base, toute transformation $\mathbf{A}$ possède une matrice A. Vu qu'aux opérations sur les transformations correspon-

dent les mêmes opérations sur leurs matrices, la transformation $p(\mathbf{A})$ sera définie par la matrice

$$p(A) = \alpha_0 E + \alpha_1 A + \alpha_2 A^2 + \dots + \alpha_m A^m.$$

Les expressions de cette forme sont appelées *polynômes matriciels*.

L'addition, la soustraction et la multiplication des matrices possèdent les mêmes propriétés principales que les opérations sur les nombres, à l'exception de la commutativité de la multiplication. Vu que pour déduire les propriétés des polynômes numériques on n'est pas obligé de diviser par une variable indépendante, le besoin d'inverser les matrices ne se présente pas. Notons que pour multiplier les polynômes matriciels, il s'avère nécessaire de multiplier les puissances d'une même matrice, lesquelles sont commutables : $A^i A^s = A^s A^i = A^{i+s}$.

Il s'ensuit que les opérations algébriques sur les polynômes matriciels d'une même matrice A ou de matrices commutables sont douées des mêmes propriétés que les opérations sur les polynômes numériques. Il va de soi que cette affirmation s'étend aux polynômes de transformations.

Rappelons que le noyau d'une application (en particulier, d'une transformation) est l'ensemble des vecteurs dont l'image est le vecteur nul. Le noyau de la transformation $\mathbf{A}$ est un sous-espace vectoriel noté Ker $\mathbf{A}$. L'ensemble des valeurs $\mathbf{A}(\mathscr{L}_n)$ de la transformation $\mathbf{A}$ est noté Im $\mathbf{A}$.

Proposition 5. *Pour toutes transformations* $\mathbf{A}$ *et* $\mathbf{B}$ *on a* $\operatorname{Ker}\mathbf{A} \subseteq \operatorname{Ker}\mathbf{BA}$. *En particulier,* $\operatorname{Ker}\mathbf{A} \subseteq \operatorname{Ker}\mathbf{A}^k$, *avec* $k \geqslant 1$.

En effet, si $x \in \operatorname{Ker}\mathbf{A}$, on a $\mathbf{A}(x) = o$ et $\mathbf{BA}(x) = o$.

La proposition 5 entraîne la formule suivante : si q est un multiple commun aux polynômes $p_1, \dots, p_s$, on a

$$\operatorname{Ker} p_i(\mathbf{A}) \subseteq \operatorname{Ker} q(\mathbf{A})$$

pour tous les $i = 1, \dots, s$.

Si chacun des sous-espaces $\operatorname{Ker} p_i(\mathbf{A})$ est inclus dans $\operatorname{Ker} q(\mathbf{A})$, il en est de même de leur somme. Donc,

$$\sum_{i=1}^{s} \operatorname{Ker} p_i(\mathbf{A}) \subseteq \operatorname{Ker} q(\mathbf{A}). \quad (2)$$

Proposition 6. Si $q = \text{PPCM}\,(p_1, \dots, p_s)$, *il vient*

$$\sum_{i=1}^{s} \operatorname{Ker} p_i(\mathbf{A}) = \operatorname{Ker} q(\mathbf{A}).$$

Etant donné la formule (2), il nous reste à démontrer que chaque vecteur de $\operatorname{Ker} q(\mathbf{A})$ se décompose en une somme de vecteurs $x_1, \dots, x_s$ tels que $x_i \in \operatorname{Ker} p_i(\mathbf{A})$.

Pour le démontrer, notons h_i le quotient de la division du polynôme q par le polynôme p_i. En vertu de la proposition 4 et du corollaire de la proposition 3, il existe des polynômes $u_1, \ldots, u_s$ tels que $1 = h_1 u_1 + \ldots + h_s u_s$. Cela signifie que la transformation identique **E** se décompose en une somme de produits :

$$\mathbf{E} = u_1(\mathbf{A})h_1(\mathbf{A}) + \ldots + u_s(\mathbf{A})h_s(\mathbf{A}).$$

Appliquons les deux membres de cette égalité à un vecteur arbitraire x de $\mathrm{Ker}\, q(\mathbf{A})$. On obtient $x = x_1 + \ldots + x_s$, où $x_i = u_i(\mathbf{A})h_i(\mathbf{A})(x)$.

Démontrons que $x_i \in \mathrm{Ker}\, p_i(\mathbf{A})$. En effet, si on substitue $p_i^{\mathbf{A}}$ à $p_i(\mathbf{A})$ pour rendre les notations plus compactes, on pourra écrire

$$p_i^{\mathbf{A}}(x_i) = p_i^{\mathbf{A}} u_i^{\mathbf{A}} h_i^{\mathbf{A}}(x) = u_i^{\mathbf{A}} q^{\mathbf{A}}(x) = 0,$$

ce qui démontre la proposition.

Soit d un diviseur commun aux polynômes $p_1, \ldots, p_s$. Cela signifie que pour tout i il existe un polynôme h_i tel que $p_i = h_i d$. Donc, la proposition 5 montre que $\mathrm{Ker}\, d(\mathbf{A})$ est inclus dans chacun des sous-espaces $\mathrm{Ker}\, p_i(\mathbf{A})$ et, par suite, dans leur intersection :

$$\mathrm{Ker}\, d(\mathbf{A}) \subseteq \bigcap_{i=1}^{s} \mathrm{Ker}\, p_i(\mathbf{A}). \tag{3}$$

Proposition 7. *Si* $d = \mathrm{PGCD}\,(p_1, \ldots, p_s)$, *on a*

$$\mathrm{Ker}\, d(\mathbf{A}) = \bigcap_{i=1}^{s} \mathrm{Ker}\, p_i(\mathbf{A}).$$

En vertu de la formule (3), il ne reste qu'à démontrer que chaque vecteur de l'intersection des espaces $\mathrm{Ker}\, p_i(\mathbf{A})$ appartient à $\mathrm{Ker}\, d(\mathbf{A})$. Pour le démontrer, écrivons d sous forme de $d = u_1 p_1 + \ldots + u_s p_s$. D'où $d^{\mathbf{A}} = u_1^{\mathbf{A}} p_1^{\mathbf{A}} + \ldots + u_s^{\mathbf{A}} p_s^{\mathbf{A}}$. Appliquons les deux membres de cette égalité à un vecteur arbitraire x de l'intersection. Il vient

$$d^{\mathbf{A}}(x) = u_1^{\mathbf{A}} p_1^{\mathbf{A}}(x) + \ldots + u_s^{\mathbf{A}} p_s^{\mathbf{A}}(x) = o,$$

ce qui est équivalent à l'assertion à démontrer.

Si les polynômes $p_1, \ldots, p_s$ sont premiers entre eux, $d(\mathbf{A}) = \mathbf{E}$. Or $\mathrm{Ker}\, \mathbf{E} = o$, de sorte que la proposition 7 permet d'énoncer le

Corollaire. *Pour des polynômes premiers entre eux on a*

$$\bigcap_{i=1}^{s} \mathrm{Ker}\, p_i(\mathbf{A}) = o.$$

3. Polynôme annulateur minimal d'une transformation. La transformation **O** d'un espace vectoriel $\mathscr{L}_n$ est dite *nulle* si elle associe à chaque vec-

teur le vecteur nul. Ceci est équivalent au fait que le noyau de la transformation nulle se confond avec l'espace $\mathscr{L}_n$ tout entier.

DÉFINITION. Un polynôme non nul p sera appelé *polynôme annulateur* de la transformation $\mathbf{A}$ si $p(\mathbf{A}) = \mathbf{O}$.

Il est évident que p est le polynôme annulateur de la transformation $\mathbf{A}$ si et seulement si $\operatorname{Ker} p(\mathbf{A}) = \mathscr{L}_n$.

Quelle que soit la base de l'espace $\mathscr{L}_n$, la transformation nulle est définie par la matrice nulle O. Par suite, la matrice A de la transformation $\mathbf{A}$ vérifie l'équation $p(A) = O$ si et seulement si p est un polynôme annulateur de $\mathbf{A}$.

PROPOSITION 8. *Pour toute transformation linéaire* $\mathbf{A}$ *il existe un polynôme qui l'annule.*

DÉMONSTRATION. Soit A la matrice de la transformation $\mathbf{A}$ dans une base quelconque. Parmi les puissances entières positives de la matrice A il y en a au plus n^2 qui sont linéairement indépendantes, car en général il existe au plus n^2 matrices linéairement indépendantes d'ordre n. Soient $\alpha_0, \ldots, \alpha_m$ les coefficients d'une combinaison linéaire non triviale et égale à zéro des puissances de la matrice A. Alors,

$$\alpha_0 \mathbf{E} + \alpha_1 \mathbf{A} + \ldots + \alpha_m \mathbf{A}^m = \mathbf{O},$$

ce qui achève la démonstration.

En parlant des polynômes annulateurs, on se place dans une situation différente de celle adoptée en algèbre élémentaire. On s'y intéresse à des racines du polynôme donné, tandis qu'ici on considère l'ensemble des polynômes dont les valeurs sur la transformation donnée $\mathbf{A}$ sont nulles.

L'ensemble $\mathscr{P}$ de tous les polynômes annulateurs de la transformation donnée $\mathbf{A}$ possède les propriétés suivantes qui sont faciles à vérifier :

a) Si $g, h \in \mathscr{P}$, alors $g + h \in \mathscr{P}$.

b) Si $g \in \mathscr{P}$, alors $gh \in \mathscr{P}$ quel que soit le polynôme h.

Ce sont les mêmes propriétés que celles de l'ensemble $\mathscr{I}$ qu'on vient de définir lors de la démonstration de la proposition 3. En partant de ces propriétés, on a démontré que l'ensemble considéré contient un polynôme de puissance minimale qui divise tous les autres polynômes. Le même raisonnement aboutit à la proposition suivante.

PROPOSITION 9. *Parmi les polynômes annulateurs de la transformation* $\mathbf{A}$ *il existe un polynôme de puissance minimale qui divise tous les polynômes annulateurs de* $\mathbf{A}$. *Ce polynôme est défini à un facteur numérique près.*

DÉFINITION. Le polynôme décrit dans la proposition 9 est appelé *polynôme minimal* de la transformation $\mathbf{A}$ si son coefficient dominant est 1.

EXEMPLE. Soit la transformation **A** définie dans une base par la matrice matrice

$$A = \begin{Vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \end{Vmatrix}.$$

Démontrons que le polynôme $t^2 - 3t + 2 = (t - 1)(t - 2)$ est le polynôme minimal de la transformation **A**. En effet, c'est un polynôme annulateur, car

$$(A - E)(A - 2E) = \begin{Vmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{Vmatrix} \cdot \begin{Vmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{Vmatrix} = O.$$

S'il existait un polynôme annulateur du premier degré $t - \alpha$, la matrice A devrait vérifier l'égalité $A - \alpha E = O$, d'où on aurait $\alpha = 1$ et $\alpha = 2$.

4. Transformations nilpotentes. DÉFINITION. La transformation linéaire **B** de l'espace vectoriel $\mathscr{K}$ est dite *nilpotente* si son polynôme minimal est de la forme t^l. Le nombre l est appelé *indice de nilpotence*.

Ainsi, si l est l'indice de nilpotence, on a $\mathbf{B}^l(x) = o$ pour tout x de $\mathscr{K}$ *et il existe un vecteur* $y \in \mathscr{K}$ tel que $\mathbf{B}^{l-1}(y) \neq o$. Il va de soi que pour certains x il s'avère que $\mathbf{B}^h(x) = o$ pour $h < l$.

Notons que toutes les valeurs propres d'une transformation nilpotente sont nulles. En effet, pour un vecteur propre x on a $\mathbf{B}^l(x) = \lambda^l x = o$, d'où $\lambda = 0$. Il s'ensuit que pour une transformation nilpotente **B** tous les vecteurs propres, ainsi que le vecteur nul, constituent le noyau Ker **B**.

PROPOSITION 10. *Supposons que* **B** *est une transformation nilpotente et que x est un vecteur qui satisfait pour un h aux conditions* $\mathbf{B}^{h-1}(x) \neq o$ *et* $\mathbf{B}^h(x) = o$. *Dans ce cas, les vecteurs* x, $\mathbf{B}(x), \ldots, \mathbf{B}^{h-1}(x)$ *sont linéairement indépendants.*

Supposons que les vecteurs sont linéairement dépendants et que α_i ($i \geqslant 0$) est le premier coefficient non nul dans leur combinaison linéaire

$$\alpha_0 x + \ldots + \alpha_i \mathbf{B}^i(x) + \ldots + \alpha_{h-1} \mathbf{B}^{h-1}(x) = o. \qquad (4)$$

Il ressort de l'hypothèse que $i < h - 1$, et l'on est en droit d'étudier la transformation $\mathbf{B}^{h-i-1}$. Considérons les images par cette transformation des deux membres de l'égalité (4). On obtient $\alpha_i \mathbf{B}^{h-1}(x) = o$, d'où $\alpha_i = 0$. La contradiction obtenue démontre la proposition.

COROLLAIRE 1. *Toute famille de vecteurs* x, $\mathbf{B}(x), \ldots, \mathbf{B}^j(x)$, *où* $\mathbf{B}^j(x)$ *est un vecteur non nul, est libre car il existe une famille libre dont elle est partie.*

COROLLAIRE 2. *L'indice de nilpotence est au plus égal à la dimension de l'espace.*

DÉFINITION. On appellera *sous-espace cyclique* relativement à la transformation nilpotente $\mathbf{B}$ l'enveloppe linéaire des vecteurs x, $\mathbf{B}(x)$, ..., $\mathbf{B}^{h-1}(x)$, si $\mathbf{B}^{h-1}(x) \neq o$ et $\mathbf{B}^h(x) = o$. On dira que le sous-espace cyclique est *engendré* par le vecteur x.

En vertu de la proposition 10, les vecteurs mentionnés dans cette définition constituent une base de l'espace cyclique. On appellera cette base *base cyclique* (engendrée par le vecteur x).

PROPOSITION 11. *Soit $\mathcal{J}$ le sous-espace cyclique de dimension h engendré par le vecteur x. Alors pour $r < h$, le sous-espace $\mathbf{B}^r(\mathcal{J})$ est cyclique. Il est engendré par le vecteur $\mathbf{B}^r(x)$ et est de dimension $h - r$. Mais si $r \geqslant h$, le sous-espace $\mathbf{B}^r(\mathcal{J})$ est nul.*

DÉMONSTRATION. Soit $y = \alpha_0 x + \alpha_1 \mathbf{B}(x) + \ldots + \alpha_{h-1}\mathbf{B}^{h-1}(x)$ un vecteur quelconque de $\mathcal{J}$. Alors, en vertu de $\mathbf{B}^h(x) = o$, l'image du vecteur y est de la forme $\mathbf{B}(y) = \alpha_0\mathbf{B}(x) + \ldots + \alpha_{h-2}\mathbf{B}^{h-1}(x)$. Selon la proposition 10, les vecteurs $\mathbf{B}(x)$, ..., $\mathbf{B}^{h-1}(x)$ constituent une base de $\mathbf{B}(\mathcal{J})$. Il en ressort que $\mathbf{B}(\mathcal{J})$ est le sous-espace cyclique engendré par le vecteur $\mathbf{B}(x)$, et $\dim \mathbf{B}(\mathcal{J}) = h - 1$ si $h > 1$. En appliquant ce résultat successivement r fois, on aboutit à la conclusion recherchée.

COROLLAIRE. *Chaque sous-espace cyclique est invariant par la transformation* $\mathbf{B}$.

En effet, de la proposition 11 on tire $\mathbf{B}(\mathcal{J}) \subset \mathcal{J}$.

§ 2. Forme normale de Jordan

1. Sous-espaces de racines. On étudiera les espaces vectoriels complexes, bien qu'en partie les résultats obtenus se vérifient aussi pour les espaces réels. A proprement parler, on devra supposer que le polynôme minimal de la transformation linéaire étudiée admet la décomposition en facteurs suivante :

$$p(t) = (t - \lambda_1)^{k_1} \ldots (t - \lambda_s)^{k_s}. \quad (1)$$

Les entiers naturels k_1, ..., k_s sont ici les multiplicités des racines λ_1, ..., λ_s. Leur somme $k_1 + \ldots + k_s$ est égale au degré du polynôme minimal. On suppose que les racines λ_1, ..., λ_s sont deux à deux différentes.

Tout polynôme présente une décomposition de la forme (1) s'il est considéré sur le corps des nombres complexes. Aussi les résultats ultérieurs se vérifient-ils pour toute transformation linéaire de l'espace complexe et pour les seules transformations linéaires de l'espace réel dont les polynômes minimaux ne possèdent que des racines réelles.

DÉFINITION. On appelle *sous-espaces de racines* de la transformation **A** les sous-espaces $\mathscr{K}_i = \mathrm{Ker}(\mathbf{A} - \lambda_i \mathbf{E})^{k_i}$, $i = 1, \dots, s$, où les nombres λ_i et k_i sont définis dans la décomposition (1) pour le polynôme minimal de la transformation **A**.

PROPOSITION 1. *Les sous-espaces de racines de la transformation* **A** *sont invariants par cette transformation.*

Chaque polynôme en **A** commute avec **A**, de sorte que le noyau de tout polynôme en **A** est invariant par **A** (proposition 1, § 2, ch. XI). En particulier, cela se rapporte aux polynômes $(\mathbf{A} - \lambda_i \mathbf{E})^{k_i}$.

THÉORÈME 1. *Si* **A** *est une transformation linéaire d'un espace vectoriel complexe $\mathscr{L}_n$, alors $\mathscr{L}_n$ est la somme directe des sous-espaces de racines $\mathscr{K}_i$ de la transformation* **A**.

DÉMONSTRATION. Ecrivons, pour abréger, la décomposition (1) du polynôme minimal de la transformation **A** sous forme de $p = b_1 b_2 \dots b_s$ et désignons par g_j le quotient de p par b_j, c'est-à-dire

$$g_j = b_1 \dots b_{j-1} b_{j+1} \dots b_s.$$

Vu qu'aucune racine n'est commune aux polynômes $g_1, \dots, g_s$, ils sont premiers entre eux. On a donc, selon la proposition 4 du § 1, $p = \mathrm{PPCM}(b_1, \dots, b_s)$.

Vu que $\mathrm{Ker}\, p(\mathbf{A}) = \mathscr{L}_n$, la proposition 6 du § 1 donne

$$\mathscr{L}_n = \sum_{i=1}^{s} \mathrm{Ker}\, b_i(\mathbf{A}) = \sum_{i=1}^{s} \mathscr{K}_i. \tag{2}$$

De façon analogue on démontre que

$$\mathrm{Ker}\, g_j = \sum_{i \neq j} \mathscr{K}_i.$$

Les polynômes b_j et g_j sont premiers entre eux, de sorte que selon le corollaire de la proposition 7, § 1, on a $\mathscr{K}_j \cap \mathrm{Ker}\, g_j = o$. Ainsi, la somme (2) est directe et le théorème est démontré.

Soit $\boldsymbol{e}$ une base de l'espace $\mathscr{L}_n$ présentant la réunion de bases des espaces de racines. Il est aisé de démontrer (comp. point 2, § 4, ch. VI) que, si les sous-espaces $\mathscr{K}_i$ sont invariants, la matrice de la transformation **A** par rapport à $\boldsymbol{e}$ se décompose en *blocs diagonaux* :

$$A = \begin{Vmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_s \end{Vmatrix}, \tag{3}$$

où $A_1, \ldots, A_s$ sont des matrices carrées d'ordres $m_1, \ldots, m_s$ égaux aux dimensions des espaces de racines. De plus, $m_1 + \ldots + m_s = n$. Les matrices A_i $(i = 1, \ldots, s)$ sont les matrices des restrictions de la transformation $\mathbf{A}$ aux sous-espaces $\mathcal{K}_i$.

La matrice décomposée en blocs diagonaux (3) sera notée également diag $(A_1, \ldots, A_s)$.

Le déterminant de la matrice $C = \operatorname{diag}(C_1, \ldots, C_s)$ est égal au produit des déterminants des blocs diagonaux $C_1, \ldots, C_s$. Pour démontrer cette proposition par récurrence sur l'ordre de la matrice, il suffit de développer le déterminant de la matrice C suivant la première ligne et appliquer l'hypothèse de récurrence à chaque mineur dans le développement obtenu.

Utilisons ce résultat pour calculer le polynôme caractéristique de la transformation $\mathbf{A}$. Soit A la matrice de $\mathbf{A}$ par rapport à une base qui est la réunion de bases des sous-espaces de racines. Alors

$$q(\lambda) = \det (A - \lambda E) = \prod_{i=1}^{s} \det (A_i - \lambda E_{m_i}), \tag{4}$$

où E_{m_i} est la matrice unité d'ordre m_i.

Ainsi, le polynôme caractéristique de la transformation est égal au produit des polynômes caractéristiques qui sont ses restrictions aux sous-espaces de racines.

Considérons l'un quelconque des sous-espaces de racines $\mathcal{K}_i = \operatorname{Ker}(\mathbf{A} - \lambda_i \mathbf{E})^{k_i}$. Soient μ une valeur propre de la restriction de $\mathbf{A}$ à ce sous-espace, et x le vecteur propre associé. Considérons l'image du vecteur x par la transformation $\mathbf{A} - \lambda_i \mathbf{E}$. Il vient

$$(\mathbf{A} - \lambda_i \mathbf{E})(x) = \mathbf{A}(x) - \lambda_i x = \mu x - \lambda_i x = (\mu - \lambda_i)x,$$

d'où

$$(\mathbf{A} - \lambda_i \mathbf{E})^{k_i}(x) = (\mu - \lambda_i)^{k_i} x = o,$$

ou bien, puisque $x \neq o$,

$$\mu = \lambda_i.$$

On voit donc que les racines du polynôme caractéristique de la restriction de $\mathbf{A}$ à $\mathcal{K}_i$ se confondent avec λ_i et, par suite, ce polynôme caractéristique est de la forme

$$p_i(\lambda) = (\lambda_i - \lambda)^{m_i}.$$

D'après la formule (4) on obtient le polynôme caractéristique de $\mathbf{A}$

$$q(\lambda) = (\lambda_1 - \lambda)^{m_1} \ldots (\lambda_s - \lambda)^{m_s}, \tag{5}$$

d'où la proposition suivante :

PROPOSITION 2. *L'ensemble des racines du polynôme minimal de la transformation* **A** *coïncide avec l'ensemble des racines de son polynôme caractéristique. Les multiplicités des racines du polynôme caractéristique sont égales aux dimensions des sous-espaces de racines.*

Notons $\mathbf{B}_i$ la restriction de la transformation $\mathbf{A} - \lambda_i \mathbf{E}$ au sous-espace de racines $\mathscr{K}_i$ correspondant à la racine λ_i. Par définition du sous-espace de racines, $\mathbf{B}_i$ est alors nilpotent et l'indice de sa nilpotence ne dépasse pas la multiplicité k_i de la racine λ_i du polynôme minimal.

PROPOSITION 3. *L'indice de nilpotence l_i de la transformation $\mathbf{B}_i$ est égal à la multiplicité de la racine λ_i du polynôme minimal.*

Supposons le contraire, soit $l_i < k_i$. Considérons le polynôme

$$\tilde{p}(\xi) = (\xi - \lambda_1)^{k_1} \dots (\xi - \lambda_i)^{l_i} \dots (\xi - \lambda_s)^{k_s}$$

et l'image par la transformation $\tilde{p}(\mathbf{A})$ d'un vecteur arbitraire x de l'espace $\mathscr{L}_n$. Décomposons x en somme de vecteurs appartenant aux sous-espaces de racines : $x = x_1 + \dots + x_s$. Il vient

$$\tilde{p}^{\mathbf{A}}(x) = \tilde{p}^{\mathbf{A}}(x_1) + \dots + \tilde{p}^{\mathbf{A}}(x_s). \tag{6}$$

Pour calculer chaque terme de (6), portons à la dernière place dans $\tilde{p}(\mathbf{A})$ le facteur qui annule le vecteur x_j de ce terme. On montrera ainsi que tous les termes, dont $\tilde{p}^{\mathbf{A}}(x_i)$, sont nuls et, par suite $\tilde{p}^{\mathbf{A}}(x) = o$ pour tout x de $\mathscr{L}_n$. Or le degré de $\tilde{p}$ est strictement inférieur à celui du polynôme minimal, ce qui contredit la définition de ce dernier. La proposition est démontrée.

Il découle des propositions 2 et 3 et du corollaire 2 de la proposition 10, § 1, que les multiplicités des racines du polynôme minimal ne dépassent pas celles du polynôme caractéristique :

$$k_i \leqslant m_i.$$

Il s'ensuit que le polynôme caractéristique est divisible par le polynôme minimal, et l'on aboutit au

THÉORÈME 2 (DE CAYLEY-HAMILTON). *Le polynôme caractéristique de toute transformation est son polynôme annulateur.*

Ce théorème admet une traduction matricielle.

THÉORÈME 2M. *Toute matrice vérifie son équation caractéristique.*

Une démonstration simple du théorème de Cayley-Hamilton a été obtenue à la suite d'une longue étude des propriétés de transformations. Le lecteur désireux de connaître les différentes démonstrations de ce théorème peut consulter la littérature spéciale (voir, par exemple Maltsev [26]). Cependant, ce théorème si élégant est rarement utilisé.

2. Chaînes de Jordan. Dans les trois points suivants, on étudiera un

sous-espace de racines fixé $\mathscr{K}_i$ et la restriction $\mathbf{B}_i$ de la transformation $\mathbf{A} - \lambda_i\mathbf{E}$ à ce dernier. La valeur de i étant unique, on omettra cet indice pour abréger l'écriture. Ainsi donc, on considère un espace $\mathscr{K}$ de dimension m et une transformation nilpotente $\mathbf{B}$ de cet espace, dont l'indice de nilpotence est k.

On démontrera que l'espace $\mathscr{K}$ se décompose en une somme directe de sous-espaces cycliques relativement à $\mathbf{B}$. A cet effet, on construira des bases cycliques dont la réunion est une base dans $\mathscr{K}$.

Soient $\mathbf{B}^h(x) \neq o$, $\mathbf{B}^{h+1}(x) = o$ pour un vecteur x. Introduisons les notations suivantes pour les vecteurs de la base cyclique :

$$\mathbf{B}^h(x) = e^0, \quad \mathbf{B}^{h-1}(x) = e^1, \ldots, \mathbf{B}(x) = e^{h-1}, \quad x = e^h.$$

Il est aisé de remarquer que le vecteur e^0 est propre pour $\mathbf{B}$, car $\mathbf{B}(e^0) = \mathbf{B}^{h+1}(x) = o$.

Définition. Supposons que des vecteurs quelconques $e^1, \ldots, e^h$ et le vecteur propre e^0 satisfont aux conditions

$$\mathbf{B}(e^1) = e^0, \quad \mathbf{B}(e^2) = e^1, \ldots, \mathbf{B}(e^h) = e^{h-1}. \tag{7}$$

On dit alors qu'ils sont respectivement les premier, deuxième, etc. vecteurs *associés* à e^0. On dit aussi que $e^0, e^1, \ldots, e^h$ constituent une *chaîne de Jordan* d'origine e^0.

On constate facilement que toute base cyclique se compose d'un vecteur propre et de vecteurs qui lui sont associés. Inversement, chaque chaîne de Jordan constitue une base cyclique dont on se convainc facilement en portant les égalités (7) l'une dans l'autre.

Proposition 4. *Le vecteur e^0 possède exactement h vecteurs associés (c'est-à-dire est l'origine de la chaîne de $h + 1$ vecteurs) si et seulement si*

$$e^0 \in \operatorname{Im} \mathbf{B}^h \cap \operatorname{Ker} \mathbf{B}, \quad e^0 \notin \operatorname{Im} \mathbf{B}^{h+1}.$$

En effet, comme il a été noté à la p. 329, l'appartenance $e^0 \in \operatorname{Ker}\mathbf{B}$ est équivalente à l'assertion que le vecteur e^0 est propre. Si $e^0 \in \operatorname{Im} \mathbf{B}^h$, il existe un vecteur x tel que $e^0 = \mathbf{B}^h(x)$. La base cyclique engendrée par le vecteur x nous fournit la chaîne recherchée. La condition $e^0 \notin \operatorname{Im} \mathbf{B}^{h+1}$ signifie qu'il n'existe pas de chaîne plus longue qui commence par e^0.

L'assertion inverse se démontre de façon analogue.

3. Recherche de l'origine de la chaîne de vecteurs. Il ressort de la proposition 4 que les sous-espaces $\operatorname{Im}\mathbf{B}^h \cap \operatorname{Ker}\mathbf{B}$ pour différents h doivent jouer un rôle important dans la construction des chaînes. On désignera $\operatorname{Im} \mathbf{B}^h \cap \operatorname{Ker}\mathbf{B}$ par $\mathscr{R}^h$.

Le vecteur $\mathbf{B}^{h+1}(x)$ pouvant être représenté sous la forme de $\mathbf{B}^h(\mathbf{B}(x))$, on a les inclusions suivantes :

$$\operatorname{Im}\mathbf{B}^{k-1} \subseteq \operatorname{Im}\mathbf{B}^{k-2} \subseteq \ldots \subseteq \operatorname{Im}\mathbf{B}^2 \subseteq \operatorname{Im}\mathbf{B}. \tag{8}$$

Si k est l'indice de nilpotence, on a $\mathrm{Im}\,\mathbf{B}^{k-1} \subsetneq \mathrm{Ker}\,\mathbf{B}$, c'est-à-dire que $\mathscr{R}^{k-1} = \mathrm{Im}\,\mathbf{B}^{k-1}$. Les intersections des sous-espaces de (8) avec le sous-espace $\mathrm{Ker}\,\mathbf{B}$ sont assurément incluses l'une dans l'autre de façon identique que les sous-espaces mêmes. Donc,

$$\mathrm{Im}\,\mathbf{B}^{k-1} = \mathscr{R}^{k-1} \subsetneq \mathscr{R}^{k-2} \subsetneq \ldots \subsetneq \mathscr{R}^{2} \subsetneq \mathscr{R}^{1} \subsetneq \mathrm{Ker}\,\mathbf{B}. \qquad (9)$$

Choisissons dans l'espace des vecteurs propres $\mathrm{Ker}\,\mathbf{B}$ une base liée aux espaces de (9). En premier lieu, rapportons à cette base tous les vecteurs d'une base quelconque de $\mathscr{R}^{k-1}$. Ensuite, joignons des vecteurs de $\mathscr{R}^{k-2}$ qui sont linéairement indépendants des précédents, puis des vecteurs de $\mathscr{R}^{k-3}$, etc. Le vecteur d'un espace $\mathscr{R}^{h-1}$ ne peut être inclus dans la base que si dans $\mathscr{R}^{h}$ il n'y a aucun vecteur qui ne se décompose en une combinaison linéaire des vecteurs déjà adjoints à la base. Ainsi, pour tout h, chaque vecteur de $\mathscr{R}^{h}$ se décompose suivant les vecteurs de la base dans $\mathscr{R}^{h}$.

Le processus décrit de l'adjonction de vecteurs s'achèvera quand la base dans $\mathrm{Ker}\,\mathbf{B}$ sera construite. Parmi les derniers vecteurs qui seront inclus dans la base de $\mathrm{Ker}\,\mathbf{B}$, on pourra aussi trouver ceux qui n'appartiennent pas à $\mathrm{Im}\,\mathbf{B}$. On désignera la base ainsi construite par $e^0 = \| e_1^0, \ldots, e_\rho^0 \|$.

4. Décomposition d'un sous-espace de racines en une somme de sous-espaces cycliques. Si à chaque vecteur de la base e_0 on adjoint une chaîne de vecteurs qui lui sont associés, on obtient le système suivant de vecteurs dans l'espace $\mathscr{K}$:

$$\begin{array}{cccc} e_1^0, & e_1^1, & \ldots, & e_1^{h_1}, \\ e_2^0, & e_2^1, & \ldots, & e_2^{h_2}, \\ \multicolumn{4}{c}{\dots\dots\dots\dots\dots\dots} \\ e_\rho^0, & e_\rho^1, & \ldots, & e_\rho^{h_\rho}. \end{array} \qquad (10)$$

Proposition 5. *Le système de vecteurs propres et de vecteurs qui leur sont associés est libre si les vecteurs propres qui y appartiennent sont linéairement indépendants.*

Démontrons-le par récurrence sur le nombre de vecteurs dans le système. S'il n'y a qu'un seul vecteur, ce dernier est propre et l'assertion est évidente.

Supposons que l'assertion est vérifiée pour tous systèmes de la forme (10) comprenant au plus $N - 1$ vecteurs. Pour un système de la forme (10) qui contient N vecteurs, considérons une combinaison linéaire quelconque égale à zéro :

$$\begin{array}{l} \alpha_1^0 e_1^0 + \alpha_1^1 e_1^1 + \ldots + \alpha_1^{h_1} e_1^{h_1} + \\ + \alpha_2^0 e_2^0 + \alpha_2^1 e_2^1 + \ldots + \alpha_2^{h_2} e_2^{h_2} + \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ + \alpha_\rho^0 e_\rho^0 + \alpha_\rho^1 e_\rho^1 + \ldots + \alpha_\rho^{h_\rho} e_\rho^{h_\rho} = o. \end{array} \qquad (11)$$

Démontrons que cette combinaison linéaire est triviale. Appliquons la transformation $\mathbf{B}$ aux deux membres de l'égalité (11). Il vient, compte tenu de (7),

$$\begin{aligned} &\alpha_1^1 e_1^0 + \ldots + \alpha_1^{h_1} e_1^{h_1 - 1} + \\ &+ \alpha_2^1 e_2^0 + \ldots + \alpha_2^{h_2} e_2^{h_2 - 1} + \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &+ \alpha_\rho^1 e_\rho^0 + \ldots + \alpha_\rho^{h_\rho} e_\rho^{h_\rho - 1} = o. \end{aligned}$$

On a donc obtenu l'égalité à zéro de la combinaison linéaire de vecteurs d'un système de la forme (10), ayant un nombre de vecteurs strictement inférieur à N (la longueur de chaque chaîne a diminué de 1). Donc, tous les coefficients de la dernière combinaison linéaire sont nuls. Ceci pris en compte dans l'égalité (11), on aboutit à l'expression suivante :

$$\alpha_1^0 e_1^0 + \ldots + \alpha_\rho^0 e_\rho^0 = o.$$

Puisque les vecteurs propres sont linéairement indépendants, ces coefficients sont également nuls. La proposition est démontrée.

Proposition 6. *Tout vecteur de l'espace $\mathscr{K}$ est une combinaison linéaire des vecteurs du système* (10).

Pour faciliter la démonstration on dira que le vecteur x de $\mathscr{K}$ est de ***hauteur*** h si $x \in \operatorname{Ker} \mathbf{B}^h$. Chaque vecteur non nul présente une certaine hauteur h, $1 \leqslant h \leqslant k$, vu que $\mathbf{B}$ est nilpotent avec indice de nilpotence k. On démontrera la proposition par récurrence sur la hauteur du vecteur.

Un vecteur non nul de hauteur 1 est propre et, par construction du système e^0, se décompose suivant les vecteurs $e_1^0, \ldots, e_\rho^0$.

Supposons que l'assertion est vérifiée pour les vecteurs de hauteur h et démontrons-la pour un vecteur arbitraire x de hauteur $h + 1$. A cet effet, considérons le vecteur $x_0 = \mathbf{B}^h(x)$. Il est évident que $x_0 \in \operatorname{Ker} \mathbf{B}$, car $\mathbf{B}(\mathbf{B}^h(x)) = o$. Donc, $x_0 \in \mathscr{R}^h = \operatorname{Im} \mathbf{B}^h \cap \operatorname{Ker} \mathbf{B}$.

Le système de vecteurs e^0 est construit de manière que chaque vecteur de $\mathscr{R}^h$ se décompose suivant des vecteurs de ce système qui appartiennent à $\mathscr{R}^h$. Aussi dans le système e^0 existe-t-il des vecteurs $e_{\alpha_1}^0, \ldots, e_{\alpha_a}^0$ de $\mathscr{R}^h$ tels que

$$x_0 = \xi_1 e_{\alpha_1}^0 + \ldots + \xi_{\alpha_a} e_{\alpha_a}^0.$$

Or les vecteurs de $\mathscr{R}^h$ possèdent des vecteurs h-ièmes associés pour lesquels $e_{\alpha_1}^0 = \mathbf{B}^h(e_{\alpha_1}^h), \ldots, e_{\alpha_a}^0 = \mathbf{B}^h(e_{\alpha_a}^h)$. Ainsi donc, on a

$$\mathbf{B}^h(x) = \xi_{\alpha_1} \mathbf{B}^h(e_{\alpha_1}^h) + \ldots + \xi_{\alpha_a} \mathbf{B}^h(e_{\alpha_a}^h).$$

On peut en conclure que le vecteur $y = x - \xi_{\alpha_1} e_{\alpha_1}^h - \ldots - \xi_{\alpha_a} e_{\alpha_a}^h$ se trouve dans $\operatorname{Ker} \mathbf{B}^h$, c'est-à-dire possède la hauteur h et vérifie l'hypothèse

de récurrence. Aussi se décompose-t-il suivant les vecteurs du système (10). Il en découle que x se décompose aussi suivant les vecteurs de ce système.

Les propositions 5 et 6 montrent qu'on a construit dans l'espace $\mathscr{H}$ une base qui est la réunion de bases cycliques. Cette base est appelée *base de Jordan* de la transformation nilpotente **B** dans l'espace $\mathscr{H}$.

Etant donné que l'enveloppe linéaire de chaque chaîne de vecteurs est un espace cyclique, on obtient le théorème suivant.

THÉORÈME 3. *Un espace $\mathscr{H}$ dans lequel est définie la transformation nilpotente* **B** *se décompose en une somme directe de sous-espaces cycliques relativement à* **B**.

5. Dimensions des sous-espaces cycliques dans la somme directe. La décomposition obtenue en vertu du théorème 3 est loin d'être unique. Mais le nombre de termes de la somme directe et leurs dimensions sont définis de façon univoque. Commençons la démonstration de ce fait par la proposition simple suivante.

PROPOSITION 7. *Si un espace $\mathscr{H}$ se décompose de deux manières en somme directe de sous-espaces cycliques, le nombre de termes non nuls est le même dans les deux décompositions.*

En effet, soit

$$\mathscr{G}_{h_1} \oplus \ldots \oplus \mathscr{G}_{h_\rho} = \mathscr{G}_{l_1} \oplus \ldots \oplus \mathscr{G}_{l_\sigma}. \tag{12}$$

La somme des dimensions des sous-espaces dans le premier et le second membre de l'égalité est égale à la dimension de $\mathscr{H}$:

$$\sum_{i=1}^{\rho} h_i = \sum_{j=1}^{\sigma} l_j = m.$$

Considérons les images par la transformation **B** des deux décompositions. On obtient deux décompositions de l'espace $\mathbf{B}(\mathscr{H})$:

$$\mathbf{B}(\mathscr{G}_{h_1}) \oplus \ldots \oplus \mathbf{B}(\mathscr{G}_{h_\rho}) = \mathbf{B}(\mathscr{G}_{l_1}) \oplus \ldots \oplus \mathbf{B}(\mathscr{G}_{l_\sigma}). \tag{13}$$

La somme y sera directe, car les sous-espaces cycliques sont invariants et tout vecteur de l'intersection de leurs images doit se trouver dans leur intersection, c'est-à-dire être nul.

Pour la décomposition (13), le calcul des dimensions selon la proposition 11 du § 1 fournit $(h_1 - 1) + \ldots + (h_\rho - 1) = (l_1 - 1) + \ldots + (l_\sigma - 1)$ ou $m - \rho = m - \sigma$, d'où $\sigma = \rho$.

THÉORÈME 4. *Supposons qu'un espace $\mathscr{H}$ admet deux décompositions* (12) *en somme directe de sous-espaces cycliques par rapport à la transformation nilpotente* **B**. *Dans ce cas, si l'une des décompositions comprend t termes de dimension h, la deuxième comprend aussi t termes de dimension h.*

DÉMONSTRATION. Soit h_1 la plus petite des dimensions de sous-espaces cycliques rencontrées dans l'une des décompositions (12). Posons pour fixer les idées qu'exactement t_1 espaces de cette dimension figurent dans la décomposition du premier membre. Appliquons à deux membres de l'égalité (12) la transformation $\mathbf{B}^{h_1}$. On obtient deux décompositions de l'espace $\mathbf{B}^{h_1}(\mathcal{K})$ en somme directe d'espaces cycliques :

$$\mathbf{B}^{h_1}(\mathcal{J}_{h_1}) \oplus \ldots \oplus B^{h_1}(\mathcal{J}_{h_\rho}) = \mathbf{B}^{h_1}(\mathcal{J}_{l_1}) \oplus \ldots \oplus \mathbf{B}^{h_1}(\mathcal{J}_{l_\sigma}).$$

Dans le premier membre de l'égalité, il reste $\rho - t_1$ termes non nuls. Un même nombre doit en rester dans le second membre. Il y avait donc exactement t_1 espaces de dimension h_1 dans le second membre de la décomposition initiale.

Admettons ensuite que le théorème est démontré pour les dimensions ne dépassant pas $h - 1$ et que la décomposition du premier membre comprend t sous-espaces de dimension h. Appliquons aux deux membres de l'égalité (12) la transformation $\mathbf{B}^h$. Elle rendra nul, dans les deux membres, tous les termes dont les dimensions sont $\leqslant h$. Vu que dans l'égalité obtenue les deux membres ont le même nombre de termes, le nombre total d'espaces de dimension $\leqslant h$ est le même dans les deux décompositions (12). En vertu de l'hypothèse de récurrence, cela signifie que le second membre de l'égalité (12) comprend exactement t sous-espaces de dimension h.

6. Matrice de la transformation nilpotente dans la base de Jordan. Considérons d'abord la restriction de la transformation nilpotente $\mathbf{B}$ à un sous-espace cyclique $\mathcal{J}$ de dimension h. Si pour base dans $\mathcal{J}$ on a choisi la chaîne des vecteurs $e^0, \ldots, e^{h-1}$, alors en vertu des relations (7), la matrice de la restriction de $\mathbf{B}$ à $\mathcal{J}$ est de la forme

$$J(0) = \left\| \begin{matrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 0 \end{matrix} \right\|, \qquad (14)$$

car les colonnes de la matrice d'une transformation sont des colonnes de coordonnées des images des vecteurs de base.

Soit maintenant un espace $\mathcal{K}$ rapporté à la base (10). Etant donné que chaque chaîne qui la compose engendre un sous-espace invariant, la matrice de la transformation $\mathbf{B}$ par rapport à cette base se décompose en blocs diagonaux diag $(J_1(0), \ldots, J_\rho(0))$, où $J_\nu(0)$ est une matrice carrée (14) d'ordre h_ν $(\nu = 1, \ldots, \rho)$.

7. Théorème de Jordan. Passons maintenant de l'étude d'un seul sous-espace de racines à celle de tout l'espace $\mathcal{L}_n$. A chaque sous-espaces de racines $\mathcal{K}_i$ correspond une racine du polynôme minimal λ_i et une transforma-

tion $\mathbf{B}_i$, restriction de la transformation $\mathbf{A} - \lambda_i \mathbf{E}$ à ce sous-espace de racines.

DÉFINITION. On appelle *base de Jordan* de la transformation **A** dans l'espace $\mathcal{L}_n$ la réunion des bases de Jordan des sous-espaces de racines, construites pour les transformations $\mathbf{B}_i$ de ces sous-espaces.

Cherchons la matrice de la transformation **A** par rapport à base de Jordan. Commençons par la matrice $A_{\mathcal{Y}}$ de la restriction de **A** à un sous-espace cyclique quelconque $\mathcal{Y}$. Notons au préalable qu'on a toutes les raisons de parler de cette restriction car un sous-espace cyclique invariant par la transformation $\mathbf{B}_i$ l'est aussi par **A**.

En vertu de la définition de la transformation $\mathbf{B}_i$, sa matrice est $A_{\mathcal{Y}} - \lambda_i E$, où E est la matrice unité d'ordre h égal à la dimension de $\mathcal{Y}$. Or la matrice de la transformation $\mathbf{B}_i$ est, comme on l'a montré, de la forme (14). Donc

$$A_{\mathcal{Y}} = J(0) + \lambda_i E = \left\| \begin{matrix} \lambda_i & 1 & 0 & \dots & 0 \\ 0 & \lambda_i & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \lambda_i & 1 \\ 0 & \dots & \dots & \dots & \lambda_i \end{matrix} \right\|.$$

Cette matrice sera notée $J(\lambda_i)$. On l'appellera *bloc de Jordan* d'ordre h à valeur propre λ_i.

Il est en fait facile de noter que la matrice $J(\lambda_i)$ possède l'unique nombre caractéristique λ_i. De plus, le vecteur commençant la chaîne qui engendre le sous-espace $\mathcal{Y}$ est le vecteur propre de la restriction de **A** à $\mathbb{X}$ et, par suite, de la transformation **A**. Cela découle directement de la forme que possède la première colonne de $J(\lambda_i)$.

Chaque espace de racines est une somme directe de sous-espaces cycliques. Aussi pour la matrice A_i de la restriction de **A** au sous-espace $\mathcal{K}_i$ a-t-on

$$A_i = \operatorname{diag}\,(J_1^i(\lambda_i), \dots, J_{p_i}^i(\lambda_i)). \tag{15}$$

Les ordres des blocs sont égaux aux dimensions des sous-espaces cycliques dont la somme directe est $\mathcal{K}_i$.

La matrice A de la transformation **A** par rapport à la réunion des bases des sous-espaces de racines est égale, comme on l'a vu, à

$$\operatorname{diag}\,(A_1, \dots, A_s).$$

On peut y porter les expressions (15) pour chaque matrice A_i :

$$A = \operatorname{diag}\,(J_1^1(\lambda_1), \dots, J_{p_s}^s(\lambda_s)). \tag{16}$$

Ce résultat peut être présenté sous forme de la

Proposition 8. *La matrice de la transformation* **A** *rapportée à sa base de Jordan est une matrice diagonale par blocs. Ses blocs diagonaux sont les blocs de Jordan dont les ordres sont égaux aux dimensions des sous-espaces cycliques en lesquels se décomposent tous les sous-espaces de racines, et dont les valeurs propres sont égales aux racines correspondantes du polynôme minimal de la transformation* **A**.

Définition. La matrice de la forme (16) décrite dans la proposition 8 sera appellée *matrice de Jordan*. Si la matrice d'une transformation linéaire est rapportée à sa base de Jordan, on dit qu'elle est réduite à la *forme normale de Jordan*.

Il importe de noter que pour obtenir la forme normale de Jordan de la matrice d'une transformation il suffit de connaître les racines du polynôme caractéristique avec leurs multiplicités, ainsi que les dimensions des sous-espaces cycliques. Le polynôme caractéristique est invariant. Quant aux dimensions des sous-espaces cycliques, elles sont définies de façon univoque en vertu du théorème 4. On aboutit au théorème suivant dénommé *théorème de Jordan*.

Théorème 5. *Pour toute transformation* **A** *de l'espace vectoriel complexe* $\mathscr{L}_n$ *il existe une base par rapport à laquelle sa matrice présente une forme normale de Jordan. Ceci étant, la forme normale de Jordan de la matrice est définie par la transformation de façon univoque à l'ordre des blocs diagonaux près.*

8. Remarques et corollaires. Comme tous les théorèmes sur les transformations linéaires, le théorème de Jordan peut être formulé en termes de matrices. Les matrices A et A' sont dites *semblables* s'il existe une matrice régulière S telle que $A' = S^{-1}AS$.

Theorème 5m. *Toute matrice est semblable à une matrice de Jordan. Pour que les matrices A et A' soient semblables il faut et il suffit que les matrices de Jordan correspondantes soient égales à l'ordre des blocs diagonaux près.*

La nécessité de la condition est évidente car deux matrices semblables peuvent être interprétées comme des matrices d'une même transformation linéaire par rapport aux bases différentes. Pour démontrer la suffisance, remarquons que les matrices J et J' qui se distinguent par l'ordre des blocs diagonaux sont semblables. En effet, la matrice correspondante S est la matrice du changement de base se réduisant à la permutation des vecteurs de base. Aussi a-t-on $A = P^{-1}JP$, $A' = Q^{-1}J'Q$ et $J' = S^{-1}JS$, d'où $A' = Q^{-1}S^{-1}PAP^{-1}SQ$.

La forme normale de Jordan n'est pas l'unique forme normale à laquelle on peut réduire la matrice d'une transformation linéaire. Dans la littérature spéciale (voir, par exemple, Maltsev [26]) le lecteur trouvera

d'autres formes normales de la matrice d'une transformation linéaire, en particulier la forme normale à laquelle se réduit la matrice de toute transformation linéaire dans l'espace vectoriel réel.

Il découle de la formule (9) que la longueur maximale de la chaîne dans un espace de racines $\mathcal{K}_i$ est égale à l'indice de nilpotence de la transformation correspondante $\mathbf{B}_i$, qui coïncide avec la multiplicité de la racine λ_i du polynôme minimal. D'où la proposition suivante.

PROPOSITION 9. *L'ordre maximal du bloc à valeur propre* λ_i *est égal à la multiplicité de la racine* λ_i *du polynôme minimal.*

On sait bien que toute transformation linéaire n'admet pas de base dans laquelle sa matrice est de forme diagonale. Le théorème de Jordan nous indique la forme la plus simple à laquelle se réduit la matrice de chaque transformation linéaire dans un espace complexe. Elle diffère de la forme diagonale par le fait qu'elle contient des unités situées immédiatement au-dessus de la diagonale principale. La matrice diagonale est un cas particulier de la matrice de Jordan.

DÉFINITION. On dira qu'une transformation linéaire est *de structure simple* (ou *à spectre simple*) si son polynôme minimal n'a pas de racines multiples.

La proposition 9 entraîne la proposition suivante.

PROPOSITION 10. *La transformation linéaire possède dans une certaine base une matrice diagonale si et seulement si cette transformation est de structure simple.*

A propos de cette assertion il convient de se rappeler l'exemple de la p. 329.

La matrice de Jordan peut être représentée sous forme de la somme d'une matrice diagonale et d'une matrice diagonale par blocs

$$B = \operatorname{diag}\,(J_1^1(0), \ldots, J_{p_s}^s(0)).$$

Démontrons que $B^r = O$ si r est l'ordre maximal du bloc. En effet, la puissance r-ième de toute matrice diagonale par blocs diag $(C_1, \ldots, C_s)$ est la matrice diag $(C_1^r, \ldots, C_s^r)$. On peut s'en assurer d'après la définition du produit des matrices. Les blocs diagonaux sont des matrices de transformations nilpotentes, de sorte que toutes leurs puissances n-ièmes s'annulent si n est supérieur ou égal à leur ordre. Ainsi, la matrice B est une matrice de la transformation nilpotente.

PROPOSITION 11. *Toute transformation linéaire d'un espace vectoriel complexe peut être représentée sous forme de la somme d'une transformation de structure simple et d'une transformation nilpotente, qui commutent entre elles.*

L'existence d'une telle représentation découle directement des raisonnements précédents. Pour démontrer la commutativité, remarquons que la matrice diagonale dont il s'agit plus haut peut être assimilée à une matrice diagonale par blocs. Ceci étant, chacun de ses blocs ne diffère de la matrice unité de même ordre que par un facteur et, par suite, commute avec le bloc correspondant de la matrice nilpotente.

9. Construction d'une base de Jordan. La démonstration donnée du théorème de Jordan est constructive, c'est-à-dire renferme un procédé de construction de la base de Jordan. Néanmoins, on a toutes les raisons de s'arrêter encore une fois sur les opérations qu'on doit accomplir pour construire la base de Jordan d'une transformation linéaire définie dans une base par la matrice A. Ce faisant, on négligera le côté du problème se rapportant aux calculs.

En particulier, on supposera qu'on est en mesure de calculer les racines du polynôme caractéristique de la matrice A et leurs multiplicités. C'est justement à partir de cela que doit débuter la construction de la base de Jordan. La suite d'opérations décrite ci-dessous doit être appliquée à chacune des racines du polynôme caractéristique.

Pour la racine λ^* on considère la matrice $A - \lambda^* E$. Si $\mathrm{Rg}\,(A - \lambda^* E) > n - m$, où m est la multiplicité de la racine λ^*, on calcule le carré, le cube, etc. de la matrice $A - \lambda^* E$, tant qu'on n'obtienne la puissance k-ième pour laquelle $\mathrm{Rg}\,(A - \lambda^* E)^k = n - m$. Le nombre k est la multiplicité de la racine λ^* du polynôme minimal. (Si $\mathrm{Rg}(A - \lambda^* E) = n - m$, la racine du polynôme minimal est simple.) En effet, si la matrice A est de la forme de Jordan, les seuls blocs nilpotents dans la matrice $A - \lambda^* E$ sont les blocs correspondant à la racine λ^*. Les autres blocs sont des matrices régulières. Dans la matrice $(A - \lambda^* E)^k$, les blocs correspondant à λ^* s'annulent et le rang de $(A - \lambda^* E)^k$ devient égal à $n - m$. Vu que le rang de cette matrice est indépendant du choix de la base, la condition imposée sur le rang de $(A - \lambda^* E)^k$ est la même dans toute base.

Considérons les colonnes qui renferment le mineur principal de la matrice $(A - \lambda^* E)^{k-1}$. Leur enveloppe linéaire est $\mathrm{Im}(\mathbf{A} - \lambda^* \mathbf{E})^{k-1}$. On ne connaît pas le sous-espace de racines mais on peut obtenir le sous-espace $\mathscr{R}^{k-1}$ (voir formule (9)) si l'on trouve l'intersection de $\mathrm{Im}(\mathbf{A} - \lambda^* \mathbf{E})^{k-1}$ avec le sous-espace $\mathscr{S}(\lambda^*)$ des vecteurs propres associés à la racine λ^*. Considérons un système maximal de vecteurs linéairement indépendants dans $\mathscr{R}^{k-1}$. Ce sont les vecteurs propres qui commencent les chaînes de longueur maximale k. Si le nombre total de vecteurs dans ces chaînes est strictement inférieur à m, on considère l'intersection $\mathscr{R}^{k-2}$. $\mathrm{Im}(\mathbf{A} - \lambda^* \mathbf{E})^{k-2} \cap \mathbf{S}(\lambda^*)$. Complétons les vecteurs propres déjà choisis jusqu'à la base dans $\mathscr{R}^{k-2}$. Les vecteurs propres ajoutés sont les origines des

chaînes de longueur $k - 1$. Si la longueur totale de toutes les chaînes est encore strictement inférieure à m, on continue la recherche de nouveaux vecteurs propres dans les espaces $\mathscr{R}^{k-3}$, ..., jusqu'à ce que la longueur totale de toutes les chaînes soit égale à m.

Il n'est pas nécessaire de trouver tous les vecteurs pour calculer la longueur de la chaîne, mais on pourra les obtenir dès qu'on aura choisi un vecteur propre e^0 qui commence la chaîne. Les vecteurs $e^1, \dots, e^h$ associés à e^0 s'obtiennent par résolution des systèmes linéaires par rapport aux colonnes de coordonnées de ces vecteurs :

$$(A - \lambda^* E)e^l = e^{l-1}, \quad l = 1, \dots, h,$$

dont le premier système contient le vecteur propre e^0. Il convient de souligner que pour les résoudre il suffit de trouver une seule solution de chaque système.

Il y a exactement h systèmes de cette forme qui sont compatibles si $e^0 \in \mathscr{R}^{h-1}$ et $e^0 \in \mathscr{R}^h$.

On a décrit ici la construction des vecteurs de base de Jordan dans un sous-espace de racines. La réunion de toutes ces bases donne une base de Jordan dans $\mathscr{L}_n$.

La matrice de Jordan d'une transformation peut être écrite aussitôt que sont connues les longueurs de toutes les chaînes correspondant à chaque racine du polynôme caractéristique.

§ 3. Fonctions de matrices

1. Introduction. En accord avec la définition générale de la fonction, on appelle fonction définie sur un ensemble $\mathscr{S}$ de matrices carrées d'ordre n à valeurs dans un ensemble $\mathscr{P}$ l'application qui à chaque matrice de l'ensemble $\mathscr{S}$ associe un élément unique de l'ensemble $\mathscr{P}$. En particulier, on s'intéressera aux fonctions dont les valeurs sont aussi des matrices carrées de même ordre. Toutefois, sans restrictions supplémentaires, on n'arrive pas à saisir dans une définition aussi large que n^2 nombres composant la matrice sont ordonnés de façon spéciale et que les opérations algébriques sont définies sur les matrices. Au fond, on pourra dire autant de cette fonction que de tout n^2-uple de fonctions de n^2 variables.

On fournira une définition plus restrictive, permettant de tenir compte de la spécificité de l'argument matriciel. Cela permet de définir pour des matrices les fonctions élémentaires telles que les fonctions exponentielle, puissance, logarithmique, trigonométrique, etc. Un exemple de ce genre a déjà été rencontré. En se servant des opérations algébriques sur les matrices, on a porté une matrice dans le polynôme comme valeur de la variable indépendante. La matrice obtenue était prise pour la valeur du polynôme

sur la matrice initiale. Chacune des fonctions élémentaires mentionnées plus haut se décompose dans un certain domaine en une série entière. En ajoutant aux opérations algébriques sur les matrices l'opération de passage à la limite, on peut définir la somme de la série entière matricielle. Les fonctions qui se laissent développées en séries entières de matrices feront justement l'objet de notre étude.

Pour assimiler ce paragraphe, le lecteur a besoin de savoir les faits élémentaires sur les séries entières d'une variable complexe.

2. Fonctions régulières de matrices. Soient données une suite numérique complexe $\{\alpha_k\}$ et une matrice carrée A. La somme formelle

$$\alpha_0 E + \alpha_1 A + \alpha_2 A^2 + \ldots$$

ou bien

$$\sum_{k=0}^{\infty} \alpha_k A^k \tag{1}$$

est appelée *série entière par rapport à la matrice A*. Les sommes finies de la forme

$$\sum_{k=0}^{N} \alpha_k A^k$$

sont appelées *sommes partielles* de la série (1).

Choisissons dans l'espace des matrices carrées d'ordre n une norme matricielle, par exemple

$$\|A\| = n \max_{i,j} |a_{ij}|. \tag{2}$$

La série (1) est dite *convergente* vers la matrice F si la suite des sommes partielles de cette série converge en norme choisie vers F, c'est-à-dire que pour tout nombre $\varepsilon > 0$ il existe un $N_0(\varepsilon)$ tel que pour tous les $N > N_0(\varepsilon)$ est vérifiée l'inégalité

$$\left\| F - \sum_{k=0}^{N} \alpha_k A^k \right\| < \varepsilon.$$

Si la série (1) converge vers la matrice F, on dira que F est la *somme* de la série et on écrira

$$F = \sum_{k=0}^{\infty} \alpha_k A.$$

La définition de la somme d'une série entière s'étend facilement aux

séries de la forme

$$\sum_{k=0}^{\infty} \alpha_k (A - A_0)^k,$$

où A_0 est une matrice fixée. Grâce à la facilité de passage des séries de cette forme aux séries de la forme (1) et inversement, on étudiera essentiellement les séries de la forme (1).

On a démontré au § 3 du ch. XI que la convergence d'une suite de matrices en norme quelconque est équivalente à la convergence en éléments. Aussi peut-on considérer une série entière matricielle comme n^2 séries numériques.

Chaque somme partielle est une matrice qui est égale à la valeur du polynôme à coefficients $\alpha_0, \ldots, \alpha_N$ sur la matrice A. Il va de soi que la somme de la série, si cette série est convergente, dépend aussi de A.

DÉFINITION. Soient $\mathscr{S}$ *et* $\mathscr{P}$ deux ensembles de matrices carrées complexes d'ordre n. La fonction sur l'ensemble $\mathscr{S}$ *à valeurs dans l'ensemble* $\mathscr{P}$ *est dite régulière* s'il existe une série entière convergeant sur $\mathscr{S}$ telle que pour chacune des matrices A de $\mathscr{S}$ est vérifié

$$f(A) = \sum_{k=0}^{\infty} \alpha_k (A - A_0)^k.$$

Une propriété importante des fonctions régulières de matrices est exprimée par la proposition suivante.

PROPOSITION 1. *Soit*

$$f(A) = \sum_{k=0}^{\infty} \alpha_k A^k.$$

Dans ce cas, pour toute matrice régulière S on a l'égalité

$$S^{-1} f(A) S = \sum_{k=0}^{\infty} \alpha_k (S^{-1} A S)^k.$$

DÉMONSTRATION. $(S^{-1}AS)^2 = S^{-1}ASS^{-1}AS = S^{-1}A^2S$. On vérifie aisément par récurrence que pour tout k on a $(S^{-1}AS)^k = S^{-1}A^kS$. D'où on obtient pour les sommes partielles de la série

$$S^{-1}\left(\sum_{k=0}^{N} \alpha_k A^k\right)S = \sum_{k=0}^{N} \alpha_k (S^{-1}AS)^k.$$

La proposition sera démontrée si on démontre que pour tout facteur à droite constant on a $\lim_{N\to\infty} (P_N S) = (\lim_{N\to\infty} P_N)S$, et qu'on démontre l'égalité analogue pour tout facteur à gauche.

La démonstration de ces égalités est absolument analogue à la démonstration connue de l'analyse. Soit $\lim_{N \to \infty} P_N = F$. Majorons la norme de la différence $P_N S - FS$:

$$\| P_n S - FS \| \leqslant \| P_N - F \| \cdot \| S \|.$$

On voit que cette norme sera inférieure à ε pour $\| P_n - F \| < \varepsilon' = \varepsilon / \| S \|$. Or on a par définition $\| P_N - F \| < \varepsilon'$ pour tous les $N > N_0(\varepsilon')$. Pour les facteurs à gauche la démonstration est presque la même.

Considérons un espace vectoriel $\mathcal{L}_n$ rapporté à une base $\boldsymbol{e}$. Soit **A** une transformation linéaire de l'espace $\mathcal{L}_n$ définie par la matrice A dans la base donnée. Si A' est la matrice de **A** par rapport à la base $\boldsymbol{e}' = \boldsymbol{e}S$, il ressort de la proposition 1 que $f(A') = S^{-1} f(A) S$. Donc, les matrices $f(A')$ et $f(A)$ définissent une même transformation linéaire quelle que soit la base $\boldsymbol{e}'$. On peut interpréter la transformation de matrice $f(A)$ comme une *valeur de la fonction f sur la transformation* **A** et la noter $f(\mathbf{A})$.

3. Etude de la convergence des séries entières matricielles. Dans l'étude des séries entières de matrices, la proposition 1 nous offre la possibilité de les simplifier par substitution de la matrice $S^{-1}AS$ à l'argument matriciel A.

PROPOSITION 2. *Soit* $A = \mathrm{diag}\,(A_1, \ldots, A_s)$. *Alors pour toute fonction régulière f définie sur la matrice A,*

$$f(A) = \mathrm{diag}\,(f(A_1), \ldots, f(A_s)).$$

En effet, comme il a été montré à la p. 341, toute puissance de la matrice diagonale par blocs est une matrice de même type qui contient les puissances des blocs diagonaux. Vu que l'addition et la multiplication par un nombre sont définies pour les matrices comme opérations sur leurs éléments, on a pour tout polynôme p

$$p(A) = \mathrm{diag}\,(p(A_1), \ldots, p(A_s)).$$

Ensuite, puisque la norme (2) de la matrice est supérieure à celle de l'un quelconque de ses blocs, les inégalités

$$\| p_N(A_i) - f(A_i) \| < \varepsilon$$

se vérifient à partir au moins du même numéro que l'inégalité

$$\| p_N(A) - f(A) \| < \varepsilon,$$

ce qui achève la démonstration de la proposition.

Etant donné une matrice carrée A d'ordre n, on peut lui faire correspondre une transformation linéaire **A** définie par la matrice A dans une base $\boldsymbol{e}$. Soit maintenant S la matrice de passage de la base $\boldsymbol{e}$ à la base qui est la réunion de bases des sous-espaces de racines de la transformation **A**.

Alors, comme on sait,

$$A' = S^{-1}AS = \text{diag}\,(A_1, \ldots, A_s),$$

où $A_1, \ldots, A_s$ sont les matrices des restrictions de **A** aux sous-espaces de racines. En vertu de la proposition 2, on obtient maintenant

$$f(A) = Sf(A')S^{-1} = S\,\text{diag}\,(f(A_1), \ldots, f(A_s))S^{-1}. \tag{3}$$

Ce résultat peut être formulé ainsi :

PROPOSITION 3. *La fonction régulière f est définie sur la matrice A de la transformation linéaire* **A** *si et seulement si elle définie sur les matrices A_i des restrictions de* **A** *à ses sous-espaces de racines. La valeur $f(A)$ peut être obtenue par la formule* (3).

Considérons maintenant un sous-espace de racines $\mathscr{K}$ de dimension m_i, correspondant à la racine λ_i. On a vu que la transformation linéaire $\mathbf{B}_i$, qui est la restriction de la transformation $\mathbf{A} - \lambda_i \mathbf{E}$ à $\mathscr{K}_i$, est nilpotente et son indice de nilpotence k_i est égal à la multiplicité de la racine λ_i du polynôme minimal. La transformation $\mathbf{B}_i$ est définie par la matrice $B_i = A_i - \lambda_i E_{m_i}$, où E_{m_i} est la matrice unité d'ordre m_i. Ainsi, la matrice A_i peut être représentée sous forme de la somme

$$A_i = B_i + \lambda_i E_{m_i}.$$

La matrice E_{m_i} commute avec B_i comme avec toute autre matrice de même ordre. Aussi la puissance r-ième de la matrice A peut-elle être écrite suivant la formule du binôme de Newton. (Rappelons que pour des matrices commutables, les règles d'opérations sur les polynômes sont les mêmes que sur les nombres.) Si C_r^j sont des coefficients binomiaux, on a

$$A_i^r = (B_i + \lambda E_{m_i})^r = \sum_{j=0}^{r} C_r^j B_i^j \lambda_i^{r-j}.$$

En vertu de la nilpotence de la matrice B_i, les seuls termes non nuls de cette décomposition sont les termes de degré $j < k_i$. Pour $r \geqslant k_i$ on a

$$A_i^r = \sum_{j=0}^{k_i-1} C_r^j \lambda_i^{r-j} B_i^j.$$

Etablissons les conditions sous lesquelles la série

$$\sum_{r=0}^{\infty} \alpha_r A_i^r \tag{4}$$

est convergente. Toute somme partielle de cette série ne contient aucune

puissance de B_i dont l'exposant est strictement supérieur à $k_i - 1$. On peut réunir les termes qui contiennent une même puissance de B_i en k_i groupes :

$$\sum_{r=0}^{N} \alpha_r A_i^r = \sum_{j=0}^{k_i-1} v_j^N B_i^j,$$

où

$$v_j^N = \sum_{r=j}^{N} \alpha_r C_r^j \lambda_i^{r-j}.$$

En se rappelant l'expression du coefficient binomial C_r^j, on trouve

$$v_j^N = \frac{1}{j!} \sum_{r=j}^{N} \alpha_r r(r-1) \dots (r-j+1) \lambda_i^{r-j}.$$

Considérons maintenant la série entière scalaire d'une variable complexe :

$$f(\xi) = \sum_{r=0}^{\infty} \alpha_r \xi^r. \tag{5}$$

Admettons que le point λ_i se trouve à l'intérieur du disque de convergence de cette série. Alors comme on le sait, on peut dériver la série terme à terme au point λ_i autant de fois que l'on veut. La dérivée d'ordre j de la fonction $f(\xi)$ au point λ_i prend la forme

$$f^{(j)}(\lambda_i) = \sum_{r=j}^{\infty} \alpha_r r(r-1) \dots (r-j+1) \lambda_i^{r-j}.$$

Le nombre v_j^N ne diffère que par un facteur de la somme partielle d'ordre N de la série de $f^{(j)}(\lambda_i)$. Donc,

$$v_j^N \to \frac{1}{j!} f^{(j)}(\lambda_i) \quad \text{pour} \quad N \to \infty.$$

Il s'ensuit que

$$\sum_{r=0}^{N} \alpha_r A_i^r = \sum_{j=0}^{k_i-1} v_j^N B_i^j \to \sum_{j=0}^{k_i-1} \frac{1}{j!} f^{(j)}(\lambda_i) B_i^j. \tag{6}$$

En effet, si on désigne $\frac{1}{j!} f^{(j)}(\lambda_i)$ par v_j, on obtient

$$\sum_{j=0}^{k_i-1} v_j^N B_i^j - \sum_{j=0}^{k_i-1} v_j B_i^j \leqslant \sum_j |v_j^N - v_j| \, \| B_i^j \| \leqslant$$

$$\leqslant \max_j \| B_i^j \| \sum_j |v_j^N - v_j|. \tag{7}$$

Mais pour chacune des k_i suites $\{v_j^N\}$ il existe pour tout $\varepsilon' > 0$ un $N_j^0(\varepsilon')$ à partir duquel on a $|v_j^N - v_j| < \varepsilon'$. Pour que le dernier membre de l'inégalité (7) soit inférieur à ε il suffit que soit vérifiée l'inégalité $N > \max_j N_j^0(\varepsilon')$, où $\varepsilon' = \varepsilon/(\max_j \|B_i^j\| \cdot k_i)$.

En définitive, on a

$$\sum_{r=0}^{\infty} \alpha_r A_i^r = \sum_{j=0}^{k_i-1} \frac{1}{j!} f^{(j)}(\lambda_i) B_i^j$$

ou, en substituant $A_i - \lambda_i E$ à B_i,

$$f(A_i) = \sum_{j=0}^{k_i-1} \frac{1}{j!} f^{(j)}(\lambda_i)(A_i - \lambda_i E)^j.$$

La relation (6) montre que pour la convergence de la série (4) il suffit que le nombre λ_i se trouve à l'intérieur du disque de convergence de la série (5). Mais si λ_i se trouve à l'extérieur du disque de convergence, même la suite $\{v_0^N\}$ n'a pas de limite. Aussi pour la convergence de la série (4) faut-il que λ_i se trouve à l'intérieur du disque de convergence ou sur sa frontière.

Profitons de la proposition 3 et de la formule (3) pour passer de la matrice A_i à la matrice de départ A. On aboutira alors au théorème suivant.

THÉORÈME 1. *Pour qu'une fonction régulière f définie par la série* (5) *soit définie sur la matrice A il suffit que les racines du polynôme caractéristique de la matrice A se trouvent à l'intérieur du disque de convergence de la série* (5). *Pour que la condition soit nécessaire il faut que ces racines appartiennent à la fermeture du disque de convergence. Dans ce cas, si S est la matrice de passage à la base située dans les sous-espaces de racines de la transformation associée à A et*

$$A = S \operatorname{diag}(A_1, \ldots, A_s) S^{-1},$$

on a

$$f(A) = S \operatorname{diag} \left\{ \sum_{j=0}^{k_1-1} \frac{1}{j!} f^{(j)}(\lambda_1)(A_1 - \lambda_1 E)^j, \ldots \right.$$
$$\left. \ldots, \sum_{j=0}^{k_s-1} \frac{1}{j!} f^{(j)}(\lambda_s)(A_s - \lambda_s E)^j \right\} S^{-1}. \qquad (8)$$

EXEMPLE 1. Appliquons la formule (8) à une matrice de structure simple. Dans ce cas, les multiplicités k_i des racines du polynôme minimal sont

égales à l'unité et dans la formule (6) il ne reste que les termes avec $B_i^0 = E_{m_i}$:

$$\sum_{r=0}^{N} \alpha_r A_i^r = \sum_{r=0}^{N} \alpha_r \lambda_i^r E_{m_i} - f(\lambda_i) E_{m_i}.$$

Ensuite, d'après la formule (3), on obtient

$$f(A) = S \operatorname{diag} (f(\lambda_1) E_{m_1}, \ldots, f(\lambda_s) E_{m_s}) S^{-1}.$$

Dans le cas particulier où A est une matrice diagonale, $f(A)$ est aussi une matrice diagonale obtenue de A par substitution du nombre $f(\lambda_i)$ à chaque élément λ_i de la diagonale. Dans le cas général, on utilise la matrice de passage S à la forme diagonale. Soit par exemple

$$A = \begin{Vmatrix} 2 & -1 \\ 0 & 1 \end{Vmatrix} = \begin{Vmatrix} 1 & 1 \\ 1 & 0 \end{Vmatrix} \cdot \begin{Vmatrix} 1 & 0 \\ 0 & 2 \end{Vmatrix} \cdot \begin{Vmatrix} 0 & 1 \\ 1 & -1 \end{Vmatrix}.$$

On obtient

$$e^A = \begin{Vmatrix} 1 & 1 \\ 1 & 0 \end{Vmatrix} \cdot \begin{Vmatrix} e & 0 \\ 0 & e^2 \end{Vmatrix} \cdot \begin{Vmatrix} 0 & 1 \\ 1 & -1 \end{Vmatrix} = \begin{Vmatrix} e^2 & e - e^2 \\ 0 & e \end{Vmatrix}.$$

EXEMPLE 2. Considérons la valeur de la fonction sur une matrice qui se compose d'un seul bloc de Jordan, par exemple d'ordre quatre :

$$A = \begin{Vmatrix} \lambda & 1 & 0 & 0 \\ 0 & \lambda & 1 & 0 \\ 0 & 0 & \lambda & 1 \\ 0 & 0 & 0 & \lambda \end{Vmatrix}.$$

Pour A, les matrices B, B^2 et B^3 sont respectivement

$$\begin{Vmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{Vmatrix}, \quad \begin{Vmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{Vmatrix}, \quad \begin{Vmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{Vmatrix}.$$

(Ici il n'y a qu'un seul sous-espace de racines, de sorte que l'indice i, qui a une seule valeur, est omis.) Par la formule (8) on obtient

$$f(A) = f(\lambda)E + f'(\lambda)B + \frac{1}{2} f''(\lambda) B^2 + \frac{1}{6} f'''(\lambda) B^3 =$$

$$= \begin{Vmatrix} f(\lambda) & f'(\lambda) & \frac{1}{2} f''(\lambda) & \frac{1}{6} f'''(\lambda) \\ 0 & f(\lambda) & f'(\lambda) & \frac{1}{2} f''(\lambda) \\ 0 & 0 & f(\lambda) & f'(\lambda) \\ 0 & 0 & 0 & f(\lambda) \end{Vmatrix}.$$

Pour formuler encore un corollaire de la formule (8), introduisons la définition suivante

DÉFINITION. On appelle *valeurs de la fonction f sur le spectre de la matrice A* les $k_1 + \ldots + k_s$ nombres

$$f(\lambda_i), f'(\lambda_i), \ldots, f^{(k_i-1)}(\lambda_i) \quad (i = 1, \ldots, s),$$

où $\lambda_1, \ldots, \lambda_s$ sont les racines du polynôme minimal de la matrice A et $k_1, \ldots, k_s$ leurs multiplicités.

PROPOSITION 4. *Si les racines du polynôme minimal de la matrice A se trouvent à l'intérieur du disque de convergence de la série scalaire qui définit la fonction f, la valeur de la fonction f sur la matrice A est définie de façon univoque par les valeurs de f sur le spectre de A.*

La proposition découle immédiatement de la formule (8). Or cette formule exprime $f(A)$ au moyen des valeurs de f sur le spectre de A d'une façon assez compliquée. Son application nous oblige de changer de base et de composer la matrice avec des blocs diagonaux. On montrera plus loin comment on peut éviter ces inconvénients.

Le symbole diag $(f(A_1), \ldots, f(A_s))$ peut être interprété comme une opération associant à l'ensemble des matrices $f(A_1), \ldots, f(A_s)$ une matrice à blocs diagonaux $f(A_1), \ldots, f(A_s)$. Si l'on passe des matrices aux transformations, l'opération diag se laisse interprétée de la façon suivante.

Soit un espace décomposé en somme directe de sous-espaces $\mathscr{K}_1, \ldots$ $\ldots, \mathscr{K}_s$. Supposons qu'une transformation $f(\mathbf{A}_i)$ est définie sur chaque $\mathscr{K}_i$. On construit une transformation $f(\mathbf{A})$ pour laquelle les $\mathscr{K}_i$ sont des sous-espaces invariants et les $f(\mathbf{A}_i)$ sont des restrictions à ces sous-espaces. Proposons-nous de réaliser cette construction indépendamment de la base. Ceci étant, les transformations $f(\mathbf{A}_i)$ des sous-espaces $\mathscr{K}_i$ seront définies non pas directement mais comme des restrictions de certaines transformations $\mathbf{F}_i$ définies sur tout l'espace. Soit donc

$$\mathbf{F}_i = \sum_{j=0}^{k_i-1} \frac{1}{j!} f^{(j)}(\lambda_i)(\mathbf{A} - \lambda_i \mathbf{E})^j. \tag{9}$$

Cette transformation est un polynôme en $\mathbf{A}$, de sorte que les sous-espaces $\mathscr{K}_i, \ldots, \mathscr{K}_s$ sont invariants par $\mathbf{F}_i$. Les restrictions de $\mathbf{F}_i$ à $\mathscr{K}_l$, avec $l \neq i$, ne nous intéressent pas, quant à la restriction de $\mathbf{F}_i$ à $\mathscr{K}_i$, c'est $f(\mathbf{A}_i)$.

Pour construire $f(\mathbf{A})$ à partir des $\mathbf{F}_i$, il nous faut une plus ample information.

4. Injection canonique et projecteur. Considérons un sous-espace $\mathscr{L}'$ de l'espace vectoriel $\mathscr{L}$. Tout vecteur de $\mathscr{L}'$ appartient à $\mathscr{L}$, et l'on peut définir une application $\mathbf{I} : \mathscr{L}' \to \mathscr{L}$ qui à chaque vecteur $x \in \mathscr{L}'$ associe le

même vecteur, considéré comme un vecteur de $\mathscr{L}$. *Cette application est appelée injection canonique* de $\mathscr{L}'$ dans $\mathscr{L}$.

Si une base $\boldsymbol{e}$ dans $\mathscr{L}$ est obtenue par adjonction de vecteurs à la base $\boldsymbol{e}'$ de $\mathscr{L}'$, la matrice de l'injection canonique par rapport aux bases $\boldsymbol{e}'$ et $\boldsymbol{e}$ contient des colonnes de la matrice unité d'ordre n, dont les numéros sont ceux des vecteurs de la base $\boldsymbol{e}'$ qu'ils ont dans la base $\boldsymbol{e}$.

Admettons maintenant que l'espace $\mathscr{L}$ est décomposé en somme directe de sous-espaces $\mathscr{K}_1, \ldots, \mathscr{K}_s$. Cela signifie que chaque vecteur x de $\mathscr{L}$ se décompose de façon univoque en somme de la forme $x = x_1 + \ldots + x_s$, où $x_i \in \mathscr{K}_i$ pour tous les i. Cette décomposition permet de définir pour chaque sous-espace $\mathscr{K}_i$ une application $\mathbf{P}_i : \mathscr{L} \to \mathscr{K}_i$ telle que $\mathbf{P}_i(x) = x_i$.

L'application $\mathbf{P}_i$ est appelée *projecteur* de $\mathscr{L}$ sur $\mathscr{K}_i$. Il faut toutefois se rappeler qu'elle est définie par toute la famille de sous-espaces $\mathscr{K}_1, \ldots, \mathscr{K}_s$ et non pas par un seul $\mathscr{K}_i$.

Si la base de $\mathscr{L}$ est la réunion de bases des sous-espaces $\mathscr{K}_1, \ldots, \mathscr{K}_s$, la matrice du projecteur $\mathbf{P}_i$ se compose des lignes de la matrice unité d'ordre n, dont les numéros sont égaux aux numéros des vecteurs de base se trouvant dans $\mathscr{K}_i$.

Supposons maintenant que les sous-espaces $\mathscr{K}_i$ sont invariants par une transformation $\mathbf{F}$ et notons $\mathbf{C}_i$ les restrictions de $\mathbf{F}$ à $\mathscr{K}_i$. On a dans ce cas l'égalité

$$\mathbf{C}_i\mathbf{P}_i = \mathbf{P}_i\mathbf{F}. \tag{10}$$

En effet, pour tout x de $\mathscr{L}$ on a

$$\mathbf{F}(x) = \sum \mathbf{F}(x_i) = \sum \mathbf{C}_i(x_i),$$

avec $x_i \in \mathscr{K}_i$ et $\mathbf{C}_i(x_i) \in \mathscr{K}_i$. Donc, $\mathbf{P}_i\mathbf{F}(x) = \mathbf{C}_i(x_i)$.

D'autre part, il est évident que $\mathbf{C}_i\mathbf{P}_i(x) = \mathbf{C}_i(x_i)$. La formule (10) est ainsi démontrée.

5. Décomposition spectrale. Revenons au calcul de $f(\mathbf{A})$. On est maintenant en mesure de montrer que

$$f(\mathbf{A}) = \sum_{i=1}^{s} \mathbf{I}_i\mathbf{P}_i\mathbf{F}_i, \tag{11}$$

où $\mathbf{P}_i$ est le projecteur sur le sous-espace de racines $\mathscr{K}_i$, $\mathbf{I}_i$ l'injection canonique définie sur $\mathscr{K}_i$, et $\mathbf{F}_i$ se définit par la formule (9). En effet, selon (10), il vient

$$\sum_{i=1}^{s} \mathbf{I}_i\mathbf{P}_i\mathbf{F}_i = \sum_{i=1}^{s} \mathbf{I}_i f(\mathbf{A}_i)\mathbf{P}_i.$$

Considérons la transformation se trouvant dans le second membre de l'égalité, et déterminons sa matrice par rapport à la base qui est la réunion de

bases des sous-espaces de racines. Il faut pour cela multiplier, pour chaque i, les matrices I_i, $f(A_i)$ et P_i de types respectifs (n, m), (m_i, m_i) et (m_i, n). Compte tenu de la forme des matrices I_i et P_i qui a été indiquée au point 4, on trouve que $I_i f(A_i) P_i$ est une matrice carrée d'ordre n dont un seul bloc sur la diagonale principale est différent de zéro. Ce bloc se dispose dans les colonnes qui correspondent aux vecteurs de base de $\mathscr{K}_i$ et vaut $f(A_i)$. Schématiquement cette multiplication peut être représentée ainsi :

Ensuite, en additionnant les matrices obtenues, on trouve que la transformation

$$\sum_{i=1}^{s} \mathbf{I}_i f(\mathbf{A}_i)\mathbf{P}_i$$

se définit dans la base envisagée par la matrice diag $(f(A_1), \ldots, f(A_s))$. Donc, elle est égale à $f(\mathbf{A})$, ce qu'il fallait démontrer.

En tenant compte de la formule (9) et en portant le produit $\mathbf{I}_i\mathbf{P}_i$ sous le symbole de sommation, on peut écrire (11) sous la forme

$$f(\mathbf{A}) = \sum_{i=1}^{s} \sum_{j=0}^{k_i-1} \frac{1}{j!} f^{(j)}(\lambda_i)\mathbf{I}_i\mathbf{P}_i(\mathbf{A} - \lambda_i\mathbf{E})^j. \qquad (12)$$

Il découle de la formule (12) qu'on vient de démontrer que quelle que soit la base, on a

$$f(A) = \sum_{i=1}^{s} \sum_{j=0}^{k_i-1} I_i P_i \frac{1}{j!} f^{(j)}(\lambda_i)(A - \lambda_i E)^j. \qquad (13)$$

Pour rendre le résultat plus explicite, introduisons la

DÉFINITION. Les matrices

$$Z_{ij} = I_i P_i \frac{1}{(j-1)!}(A - \lambda_i E)^{j-1}, \quad i = 1, \ldots, s, \ j = 1, \ldots, k_i, \qquad (14)$$

sont appelées *matrices composantes* de la matrice A. En particulier, $Z_{i1} = I_i P_i$.

Maintenant l'égalité (13) équivaut au théorème suivant.

THÉORÈME 2. *La matrice $f(A)$ est la combinaison linéaire suivante des matrices composantes* :

$$f(A) = \sum_{i=1}^{s} \sum_{j=1}^{k_i} f^{(j-1)}(\lambda_i) Z_{ij}. \tag{15}$$

L'égalité (15) est appelée *décomposition spectrale* de $f(A)$. Notons les propriétés suivantes de cette décomposition.

1. Les matrices Z_{ij} sont indépendantes de la fonction f, de sorte que pour des fonctions régulières différentes, les décompositions spectrales ne diffèrent que par les valeurs de la fonction sur le spectre de A.

2. La matrice $f(A)$ s'exprime linéairement par les valeurs de f sur le spectre de A.

3. Pour toutes les fonctions dont les valeurs sur le spectre de A coïncident, la valeur de $f(A)$ est la même.

La dernière propriété peut être utilisée pour réduire le calcul de $f(A)$ à celui d'un polynôme de A.

PROPOSITION 5. *Quels que soient la matrice A et les nombres β_{ij}, $i = 1, \ldots, s$, $j = 1, \ldots, k_s$, il existe un polynôme et un seul dont le degré est strictement inférieur au degré k du polynôme minimal et dont les valeurs sur le spectre de la matrice A sont β_{ij}.*

DÉMONSTRATION. La valeur du polynôme en un point fixé est une fonction linéaire de ses coefficients. Il en est de même des valeurs de ses dérivées de tous ordres. Par conséquent, l'exigence que le polynôme présente sur le spectre de la matrice A les valeurs mentionnées est équivalente à un système d'équations linéaires en ses coefficients. Si le degré du polynôme est $\leqslant k - 1$, le nombre d'inconnues dans le système est k. Le nombre des équations est égal à celui des valeurs sur le spectre, c'est-à-dire est aussi k. Il reste donc à démontrer que le déterminant de la matrice du système est différent de zéro. On le démontrera si on montre que le système homogène associé ne possède qu'une solution nulle.

Démontrons la dernière assertion. Considérons pour cela le polynôme qui présente sur le spectre de A les valeurs nulles. Il découle de (15) que c'est un polynôme annulateur pour la matrice A. Or son degré est strictement inférieur à celui du polynôme minimal, et il doit être nul. La proposition est démontrée.

Le polynôme décrit dans la proposition 5 est appelé *polynôme d'interpolation de Lagrange.*

6. Propriétés des matrices composantes. Démontrons la proposition qui suit.

PROPOSITION 6. *Chaque matrice composante est un polynôme de A de degré strictement inférieur à celui du polynôme minimal.*

Considérons une matrice composante arbitraire Z_{ij}. Soit h_{ij} le polynôme de Lagrange pour lequel $h_{ij}^{(j-1)}(\lambda_i) = 1$, les autres valeurs sur le spectre étant nulles. De la formule (15) il découle que $h_{ij}(A) = Z_{ij}$.

Proposition 7. *Pour toute matrice A, les matrices composantes sont linéairement indépendantes.*

Démonstration. Considérons une combinaison linéaire nulle des matrices composantes

$$\sum_{i=1}^{s} \sum_{j=1}^{k_i} \gamma_{ij} Z_{ij} = O.$$

C'est un polynôme de la matrice A, égal à zéro. Etant donné que son degré est strictement inférieur à celui du polynôme minimal, c'est un polynôme nul. Cela signifie que pour les polynômes h_{ij} de la proposition 6 est vérifiée l'égalité

$$\sum_{i=1}^{s} \sum_{j=1}^{k_i} \gamma_{ij} h_{ij} = 0. \tag{16}$$

Démontrons qu'il en résulte que $\gamma_{ij} = 0$ pour tous les i et j. En portant dans l'identité (16) le nombre λ_i, $i = 1, \ldots, s$, on obtient que $\gamma_{i1} = 0$. Ensuite, dérivons cette identité et portons λ_i dans la dérivée du premier membre. On obtient $\gamma_{i2} = 0$. On continuera à dériver et de porter λ_i jusqu'à l'ordre $k^* = \max k_i$. On obtiendra ainsi toutes les égalités exigées.

Vu que les matrices composantes sont des polynômes de A, on a la

Proposition 8. *Les matrices composantes de la matrice A commutent entre elles et avec la matrice A.*

Mentionnons quelques relations algébriques que vérifient les matrices composantes. D'abord, pour tous $i = 1, \ldots, s$, les matrices Z_{i1} sont idempotentes. Cela signifie que

$$Z_{i1}^r = Z_{i1}$$

pour tout exposant r. La démonstration est immédiate. En effet, l'image d'un vecteur x par la transformation $\mathbf{Z}_{i1}$ de matrice $Z_{i1} = I_i P_i$ est un vecteur obtenu par la projection de x sur $\mathscr{K}_i$, suivie de l'injection canonique dans $\mathscr{L}$. Il est évident qu'une succession de ces transformations ne fait pas varier le résultat.

En portant $f(\xi) = 1$ dans la décomposition (15), on vérifie que

$$\sum_{i=1}^{s} Z_{i1} = E.$$

Etant donné que le facteur à droite dans l'expression $\mathbf{I}_i \mathbf{B}_i^{j-1} \mathbf{P}_i = \mathbf{Z}_{ij}$ est

un projecteur, on a Im $\mathbf{Z}_{ij} \subseteq \mathcal{K}_i$ et $\mathcal{K}_l \subseteq$ Ker $\mathbf{Z}_{ij}$ pour $l \neq i$. D'où

$$Z_{ij} Z_{lm} = O \quad \text{pour} \quad i \neq l \quad \text{et tous les } j, m.$$

La transformation $\mathbf{Z}_{i1}$ agit sur $\mathcal{K}_i$ comme une transformation identique, de sorte que pour tous les i et j

$$Z_{i1} Z_{ij} = Z_{ij}.$$

Notons encore une égalité intéressante qu'on obtient pour $f(\xi) = \xi$

$$A = \sum_{i=1}^{s} (\lambda_i Z_{i1} + Z_{i2}).$$

Il est instructif d'étudier les matrices composantes d'une matrice A possédant la forme de Jordan. Au lieu d'une description générale encombrante, on donnera un exemple suffisamment caractéristique. Dans les matrices d'ordre six écrites plus bas, les éléments qui manquent sont égaux à zéro. Soit

$$A = \left\| \begin{matrix} 2 & & & & & \\ & 2 & 1 & & & \\ & & 2 & & & \\ & & & 3 & 1 & \\ & & & & 3 & 1 \\ & & & & & 3 \end{matrix} \right\|.$$

On a alors pour une fonction f

$$f(A) = \left\| \begin{matrix} f(2) & & & & & \\ & f(2) & f'(2) & & & \\ & & f(2) & & & \\ & & & f(3) & f'(3) & \frac{1}{2} f''(3) \\ & & & & f(3) & f'(3) \\ & & & & & f(3) \end{matrix} \right\|.$$

En vertu de l'indépendance linéaire des matrices composantes, on obtient de (15)

$$Z_{11} = \left\| \begin{matrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & 0 \end{matrix} \right\|, \qquad Z_{21} = \left\| \begin{matrix} 0 & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \end{matrix} \right\|,$$

$$Z_{12} = \begin{Vmatrix} 0 & & & & & \\ & 0 & 1 & & & \\ & & 0 & & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & 0 \end{Vmatrix}, \quad Z_{22} = \begin{Vmatrix} 0 & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & 0 & 1 & 0 \\ & & & & 0 & 1 \\ & & & & & 0 \end{Vmatrix},$$

$$Z_{23} = \begin{Vmatrix} 0 & & & & & \\ & 0 & & & & \\ & & 0 & & & \\ & & & 0 & 0 & \frac{1}{2} \\ & & & & 0 & 0 \\ & & & & & 0 \end{Vmatrix}.$$

7. Calcul des matrices composantes. On peut construire le polynôme d'interpolation de Lagrange en partant du problème d'interpolation d'Hermite. Ce problème consiste à construire pour une fonction donnée f et pour s points différents $\lambda_1, \ldots, \lambda_s$ un polynôme qui, avec ses dérivées d'ordres mentionnés $k_1, \ldots, k_s$, présente aux points $\lambda_1, \ldots, \lambda_s$ les mêmes valeurs que la fonction f. Ce polynôme de degré inférieur à $k = k_1 + \ldots + k_s$ est aussi appelé *polynôme d'interpolation d'Hermite.*

Le polynôme d'Hermite peut être obtenu d'après la formule

$$H(\xi) = \sum_{i=1}^{s} \sum_{j=1}^{k_i} f^{(j-1)}(\lambda_i) h_{ij}(\xi),$$

si sont construits les polynômes h_{ij}, $i = 1, \ldots, s$, $j = 1, \ldots, k_s$, de degrés inférieurs à k, pour lesquels

$$h_{lm}^{(j-1)}(\lambda_i) = \begin{cases} 1 \text{ pour } i = l \text{ et } j = m, \\ 0 \text{ pour } i \neq l \text{ ou } j \neq m. \end{cases} \tag{17}$$

Décrivons la construction de ces polynômes. Considérons pour chaque l le polynôme

$$u_l(\xi) = \sum_{i \neq l} \left(\frac{\xi - \lambda_l}{\lambda_i - \lambda_l} \right)^{k_i}.$$

Il est aisé de vérifier que pour tout $i \neq l$ et tout $j \leqslant k_i$

$$u_l^{(j-1)}(\lambda_i) = 0. \tag{18}$$

Cela signifie que u_l vérifie les conditions (17) aux points λ_i, $i \neq l$, mais, en général, ne les vérifie pas au point λ_l. Pour remédier à ce défaut, notons que les propriétés (18) se conservent pour tout produit par un polynôme $w(\xi)$:

$$\frac{d^{j-1}}{d\xi^{j-1}}(u_l w)\Big|_{\xi=\lambda_i} = 0, \; i \neq l, \; j \leqslant k.$$

Cela découle de la formule de Leibniz. Choisissons le polynôme w de la façon qui nous convient.

Soit $f_{lr}(\xi)$ le polynôme de Taylor obtenu par décomposition de la fonction $1/u_l$ suivant la formule de Taylor au voisinage du point λ_l à $o((\xi - \lambda_l)^r)$ près. On obtient

$$h_{lm}(\xi) = u_l(\xi)\frac{1}{(m-1)!}(\xi - \lambda_l)^{m-1} f_{l,\,k_l - m}(\xi). \tag{19}$$

Ce polynôme permet une représentation différente :

$$h_{lm}(\xi) = \frac{1}{(m-1)!}(\xi - \lambda_l)^{m-1}u_l(\xi)\left[\frac{1}{u_l} + o((\xi - \lambda_l)^{k_l - m})\right] =$$

$$= \frac{1}{(m-1)!}(\xi - \lambda_l)^{m-1} + u_l(\xi)o((\xi - \lambda_l)^{k_l - 1}).$$

Démontrons que le polynôme (19) satisfait aux conditions (17). Pour tous les $j \leqslant k_l$ on a $(u_l(\xi)o((\xi - \lambda_l)^{k_l-1}))^{(j-1)}_{\xi=\lambda_l} = 0$ selon la formule de Leibniz, vu que les $k_l - 1$ premières dérivées de la fonction $o((\xi - \lambda_l)^{k_l-1})$ au point λ_l sont nulles. Donc, les $k_l - 1$ premières dérivées du polynôme h_{lm} au point λ_l sont les mêmes que celles de $\frac{1}{(m-1)!}(\xi - \lambda_l)^{m-1}$, c'est-à-dire

$$h_{lm}^{(j-1)}(\lambda_l) = \begin{cases} 1, & j = m, \\ 0, & j \neq m. \end{cases}$$

Le polynôme u_{lm} est de degré $k - k_l$. Le produit des deux autres facteurs de (19) est de degré $\leqslant m - 1 + k_l - m$, de sorte que le degré du polynôme h_{lm} est $\leqslant k - 1$. La démonstration est ainsi achevée.

Le procédé général de calcul des matrices composantes de la matrice A consiste en la substitution de A dans les formules (19), vu que $Z_{ij} = h_{ij}(A)$ pour tous les i et j. Toutefois, il n'est pas inutile de noter un procédé plus efficace quoique moins général.

On peut porter dans la décomposition (15), au lieu de f, des polynômes quelconques, par exemple les diviseurs du polynôme minimal ou tout simplement les puissances de la matrice A. On obtient ainsi un système d'équations linéaires à coefficients numériques par rapport aux matrices Z_{ij}. Si les

polynômes ont été bien choisis, en résolvant ce système on obtient facilement les matrices composantes.

EXEMPLE. Soit

$$A = \begin{Vmatrix} 2 & -1 \\ 0 & 1 \end{Vmatrix}.$$

Etant donné que c'est une matrice d'ordre deux à spectre simple et $\lambda_1 = 1$, $\lambda_2 = 2$, la formule (15) prend la forme $f(A) = f(1)Z_{11} + f(2)Z_{21}$. Remplaçons f par les polynômes $\xi - 1$ et $\xi - 2$. Il vient immédiatement

$$Z_{21} = \begin{Vmatrix} 1 & -1 \\ 0 & 0 \end{Vmatrix}, \quad Z_{11} = \begin{Vmatrix} 0 & 1 \\ 0 & 1 \end{Vmatrix}.$$

Donc,

$$f(A) = f(1)\begin{Vmatrix} 0 & 1 \\ 0 & 1 \end{Vmatrix} + f(2)\begin{Vmatrix} 1 & -1 \\ 0 & 0 \end{Vmatrix}.$$

En particulier, comme on l'a vu à la p. 350,

$$e^A = \begin{Vmatrix} e^2 & e - e^2 \\ 0 & e \end{Vmatrix},$$

ou pour la fonction ξ^r, avec un r naturel,

$$A^r = \begin{Vmatrix} 2^r & 1 - 2^r \\ 0 & 1 \end{Vmatrix}.$$

8. Extension des identités aux matrices. Posons que la fonction régulière $g(\xi)$ est égale à zéro sur le spectre de la matrice A. Alors $g(A) = O$. Appliquons cette propriété à la fonction de la forme

$$g(\xi) = G(f_1(\xi), \ldots, f_r(\xi)) = 0,$$

où $G = 0$ est une identité reliant les fonctions régulières $f_1, \ldots, f_r$. Si la fonction $g(\xi)$ est régulière, on voit que $g(A) = O$ et par suite, l'identité se conserve aussi pour la matrice A.

Dans deux cas importants la régularité de la fonction $g(\xi)$ est hors de doute.

1) $G(\eta_1, \ldots, \eta_l)$ est un polynôme. On peut étendre aux matrices toute identité de type considéré, dont le premier membre est un polynôme de fonctions régulières. Considérons par exemple l'identité $\sin^2 \xi + \cos^2 \xi = 1$. La fonction $g(\xi) = \sin^2 \xi + \cos^2 \xi - 1 = 0$ est régulière et s'annule sur le spectre de toute matrice. Aussi pour toute matrice A a-t-on

$$\sin^2 A + \cos^2 A = E.$$

On a trouvé plus haut les matrices composantes de

$$A = \begin{Vmatrix} 2 & -1 \\ 0 & 1 \end{Vmatrix}.$$

En se servant d'elles, on obtient

$$\sin^2 A = \begin{Vmatrix} \sin^2 2 & -\sin^2 2 + \sin^2 1 \\ 0 & \sin^2 1 \end{Vmatrix}, \quad \cos^2 A = \begin{Vmatrix} \cos^2 2 & -\cos^2 2 + \cos^2 1 \\ 0 & \cos^2 1 \end{Vmatrix}.$$

Maintenant on peut se convaincre immédiatement que l'égalité est vraie pour cette matrice.

2) Soient f une fonction régulière d'une seule variable et h une branche régulière de la fonction réciproque, qui est définie dans le domaine contenant le spectre de la matrice A. Dans ce cas, la fonction $g(\xi) = h(f(\xi)) - \xi$ est régulière et l'identité $h(f(\xi)) = \xi$ se vérifie de même pour la matrice A.

A titre d'exemple, définissons la fonction $\xi^{1/2}$ comme une somme de la série

$$\xi^{1/2} = 1 + \frac{1}{2}(\xi - 1) - \frac{1}{8}(\xi - 1)^2 + \ldots$$

Vu que la fonction $(\xi^{1/2})^2 - \xi$ est régulière, la valeur

$$A^{1/2} = E + \frac{1}{2}(A - E) - \frac{1}{8}(A - E)^2 + \ldots$$

vérifie l'égalité $(A^{1/2})^2 = A$ pour toute matrice A dont les nombres caractéristiques appartiennent au disque $|\xi - 1| < 1$. Donc, $A^{1/2}$ peut être interprété comme la racine carré de A. Soit

$$A = \begin{Vmatrix} 3 & -1 \\ 0 & 2 \end{Vmatrix}.$$

Les nombres caractéristiques de cette matrice sont 3 et 2. La matrice $A - E$ est justement la matrice dont les matrices composantes ont été calculées. Aussi la somme de la série est-elle égale à

$$X = \sqrt{3}\begin{Vmatrix} 1 & -1 \\ 0 & 0 \end{Vmatrix} + \sqrt{2}\begin{Vmatrix} 0 & 1 \\ 0 & 1 \end{Vmatrix} = \begin{Vmatrix} \sqrt{3} & -\sqrt{3} + \sqrt{2} \\ 0 & \sqrt{2} \end{Vmatrix}.$$

En élevant X au carré, on obtient la matrice A.

Il convient de souligner que les identités comprenant deux variables indépendantes ne sont plus vraies pour les matrices. Cela s'explique par la non-commutativité de la multiplication dcs matrices. L'exemple classique est fourni par

$$e^A \cdot e^B \neq e^{A+B},$$

où A et B sont des matrices quelconques. Or si les matrices sont commutables, on obtient une égalité. On peut s'en convaincre en multipliant les séries entières.

9. Prolongement analytique. A propos de l'exemple, discuté plus haut, de l'extraction de la racine carrée de la matrice il faut faire deux remarques importantes. D'abord, par cette voie, on n'a obtenu qu'une des matrices X vérifiant l'équation $X^2 = A$ et possédant donc le droit d'être appelée racine carrée de A. Ensuite, on est en mesure de trouver $A^{1/2}$ pour les seules matrices dont les valeurs propres se trouvent à l'intérieur du disque de convergence, bien que la racine carrée existe aussi pour d'autres matrices.

Il va de soi que de telles remarques se rapportent non seulement à la racine carrée mais également à de nombreuses autres fonctions, par exemple au logarithme. Le lecteur versé dans la théorie des fonctions d'une variable complexe notera que les deux questions sont étroitement liées entre elles et se rapportent à la possibilité de prolongement analytique de la fonction régulière d'une matrice. Leur résolution n'a rien de compliqué mais se rapporte davantage à l'analyse qu'à l'algèbre linéaire. Aussi n'approfondira-t-on pas cet aspect de la question.

Notons seulement qu'on est en mesure de définir une fonction régulière pour les matrices dont les valeurs propres n'appartiennent pas au disque de convergence de la série définissant la fonction. Il suffit pour cela de pouvoir obtenir la valeur de la fonction sur le spectre de la matrice qui nous intéresse. Alors, en appliquant le polynôme de Lagrange, on peut déterminer la valeur de la fonction sur la matrice. Le lecteur peut trouver un meilleur exposé d'argumentation de ce procédé dans la source première, l'ouvrage de Lappo-Danilevski [23].

Donnons un exemple. La matrice

$$A = \begin{Vmatrix} 2 & -1 \\ 1 & -2 \end{Vmatrix}$$

présente les nombres caractéristiques $\sqrt{3}$ et $-\sqrt{3}$. Ils sont tels que tout disque les contenant contient aussi zéro. Il s'ensuit qu'aucun développement de la fonction $1/\xi$ en série entière ne peut converger en même temps en deux points $\sqrt{3}$ et $-\sqrt{3}$. Or la fonction y est définie et prend les valeurs $(\sqrt{3})^{-1}$ et $(-\sqrt{3})^{-1}$. Cherchons les matrices composantes de la matrice A :

$$Z_{11} = \frac{1}{2\sqrt{3}} \begin{Vmatrix} 2+\sqrt{3} & -1 \\ 1 & -2+\sqrt{3} \end{Vmatrix},$$

$$Z_{21} = -\frac{1}{2\sqrt{3}} \begin{Vmatrix} 2-\sqrt{3} & -1 \\ 1 & -2-\sqrt{3} \end{Vmatrix}.$$

Leur substitution dans l'expression $(\sqrt{3})^{-1} Z_{11} + (-\sqrt{3})^{-1} Z_{21}$ nous fournit

la matrice

$$\begin{Vmatrix} 2/3 & -1/3 \\ 1/3 & -2/3 \end{Vmatrix},$$

qui est en effet la matrice inverse de A. (Cet exemple illustre le prolongement des fonctions régulières de matrices au moyen du polynôme de Lagrange mais ne doit pas être pris pour un procédé commode d'inversion des matrices d'ordre deux.)

La décomposition spectrale, autrement dit le polynôme de Lagrange, n'est pas le seul polynôme à l'aide duquel on peut calculer les fonctions d'une matrice. Il est manifestement suffisant que les valeurs de ce polynôme sur le spectre de la matrice coïncident avec les valeurs de la fonction. Indiquons ce polynôme pour la fonction ξ^{-1}.

Selon le théorème de Cayley-Hamilton, la matrice A vérifie son équation caractéristique

$$A^n + a_1A^{n-1} + \ldots + a_{n-1}A + a_nE = O,$$

où $a_n = \det A$. Si $a_n \neq 0$, on peut écrire

$$a_n^{-1}A(A^{n-1} + a_1A^{n-2} + \ldots + a_{n-1}E) = E,$$

d'où on obtient l'expression de A^{-1} sous forme de polynôme de A.

L'utilité pratique de ce procédé d'inversion d'une matrice n'est pas grande, car le calcul des coefficients du polynôme caractéristique est une opération laborieuse. Toutefois, si le polynôme caractéristique présente un intérêt en soi, on peut en même temps trouver la matrice inverse. L'algorithme de recherche des coefficients du polynôme caractéristique et de la matrice inverse est donné dans la littérature spéciale.

10. Nombres caractéristiques de la fonction régulière. Démontrons la proposition suivante.

PROPOSITION 9. *Si f est une fonction régulière, les racines du polynôme caractéristique de la matrice $f(A)$ sont $f(\lambda_1), \ldots, f(\lambda_n)$, où $\lambda_1, \ldots, \lambda_n$ sont les racines du polynôme caractéristique de la matrice A.*

DÉMONSTRATION. En se servant de la proposition 1, on peut remplacer la matrice A par la matrice de Jordan A' telle que $A' = S^{-1}AS$. La forme de la matrice $f(J)$, où J est un bloc de Jordan d'ordre quatre, a été obtenue dans l'exemple 2 à la p. 350 et se généralise aisément aux blocs d'ordre quelconque. Pour $f(A')$ on a

$$f(A') = \operatorname{diag}(f(J_1) \ldots f(J_N)),$$

d'où l'on voit que tous les éléments dans $f(A')$ qui sont situés au-dessous de la diagonale principale sont nuls, quant aux éléments diagonaux, ce sont

les nombres $f(\lambda_1)$, ..., $f(\lambda_s)$ répétés autant de fois qu'il est nécessaire. Donc,

$$\det (f(A') - \lambda E) = (\lambda - f(\lambda_1))^{m_1} \dots (\lambda - f(\lambda_s))^{m_s},$$

ce qui est équivalent à l'assertion qu'il fallait démontrer.

Le déterminant de la matrice est égal au produit de ses nombres caractéristiques, et la trace, à leur somme (chaque nombre étant compté autant de fois qu'est sa multiplicité dans le polynôme caractéristique). D'où on obtient l'égalité suivante :

$$\det e^{A} = e^{\operatorname{tr} A}.$$

§ 4. Localisation des racines d'un polynôme caractéristique

1. Introduction. Le problème de localisation des nombres caractéristiques de la matrice consiste dans la recherche des conditions et des estimations qui permettent, d'après les éléments de la matrice, d'indiquer la position approximative des racines du polynôme caractéristique sur le plan complexe.

En principe, les éléments de la matrice permettent évidemment de déterminer les valeurs exactes des nombres caractéristiques, mais ces valeurs ne sont pas toujours nécessaires. Par ailleurs, les estimations proposées doivent être moins laborieuses que la résolution de l'équation caractéristique. En parlant de la localisation des nombres caractéristiques, il ne faut pas oublier qu'ils sont des invariants et que par suite, l'application de l'un quelconque des théorèmes de localisation à la matrice $S^{-1}AS$, au lieu de la matrice A, indique la position des mêmes nombres, mais peut fournir une estimation plus (ou moins) précise. Un choix convenable de la matrice S peut entraîner une excellente estimation jusqu'à mention précise des racines, mais ce choix est difficile.

L'intérêt qu'on porte à la localisation des nombres caractéristiques est, en premier lieu, lié à ses applications aux différents problèmes de théorie de la stabilité d'équations différentielles ordinaires et, de ce fait, à un grand nombre de problèmes pratiques. En second lieu, l'application de certaines méthodes numériques de résolution des problèmes (même de ceux, qui n'ont apparemment pas de rapport avec les nombres caractéristiques, comme par exemple les systèmes d'équations linéaires) exige une localisation, même peu exacte, des nombres caractéristiques des matrices étudiées. A plus forte raison, tout cela se rapporte au problème de recherche des vecteurs propres et des valeurs propres où une localisation préalable des nombres caractéristiques simplifie considérablement la résolution.

La localisation des nombres caractéristiques d'une matrice est intime-

ment liée à celle des racines d'un polynôme d'après ses coefficients. Mais la recherche du polynôme caractéristique présente des difficultés et ne simplifie pas le problème. Pour localiser les racines d'un polynôme d'après ses coefficients, il vaut mieux plutôt utiliser de nombreux résultats de localisation des racines d'un polynôme caractéristique d'après les éléments de la matrice. A cet effet, on peut construire une matrice pour laquelle le polynôme considéré est un polynôme caractéristique. A titre d'exemple d'une matrice de polynôme caractéristique

$$t^n + \alpha_1 t^{n-1} + \ldots + \alpha_{n-1} t + \alpha_n$$

peut servir la matrice

$$\left\|\begin{matrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 0 & 1 \\ -\alpha_n & -\alpha_{n-1} & \ldots & -\alpha_2 & -\alpha_1 \end{matrix}\right\|.$$

Elle est appelée *matrice d'accompagnement* du polynôme donné.

Notons encore un domaine qui se rattache intimement à la localisation des nombres caractéristiques. C'est la théorie des perturbations des nombres caractéristiques qui explique de quelle façon doivent varier les nombres caractéristiques et leurs multiplicités avec la faible variation (« perturbation ») des éléments de la matrice. Faute de place, on n'abordera pas ces résultats, bien que leur importance pratique soit si grande. Elle est liée aux inévitables erreurs avec lesquelles nous est présentée l'information initiale dans tout problème ainsi qu'aux erreurs d'arrondi aussi inévitables. Sous ce rapport, on fournira quelques renseignements au point 4 du § 2, ch. XIII ; l'exposé détaillé de la théorie des perturbations peut être trouvé dans les livres de Lankaster [22] et de Wilkinson [42].

Les liens mentionnés ci-dessus de la localisation des valeurs propres avec d'autres sujets en ont fait une branche importante de l'algèbre linéaire, riche en résultats variés. Dans le cadre d'un enseignement général, il est peut-être inutile d'étudier en détail tous ces résultats, aussi va-t-on présenter les théorèmes les plus simples et importants.

2. Estimations des modules des nombres caractéristiques. Introduisons la définition suivante.

DÉFINITION. Soient λ_i, $i = 1, \ldots, s$, les nombres caractéristiques de la matrice A. Le nombre $\lambda_A = \max_i |\lambda_i|$ est appelé *rayon spectral* de la matrice A.

Il va de soi que le rayon spectral est un nombre réel positif qui s'annule si et seulement si la matrice est nilpotente.

PROPOSITION 1. *Si la norme* $\|*\|$ *possède la propriété annulaire, le rayon spectral de la matrice A ne dépasse pas la norme de cette matrice* : $\lambda_A \leqslant \|A\|$.

DÉMONSTRATION. Soit λ_1 le nombre caractéristique pour lequel $|\lambda_1| = \lambda_A$ et soit $A\xi = \lambda_1\xi$, $\xi \neq 0$. Notons X la matrice carrée d'ordre n dont la première colonne est ξ et les autres colonnes sont nulles. Il est évident que $AX = \lambda_1 X$. Considérons les normes des deux membres de l'égalité. On a $|\lambda_1|\cdot\|X\| = \|AX\| \leqslant \|A\|\cdot\|X\|$. Vu que $\|X\| > 0$, on obtient l'inégalité qu'il fallait démontrer.

En appliquant la proposition 1 aux diverses normes (voir § 3, ch. XI), on obtient les majorations suivantes du module d'un nombre caractéristique arbitraire :

a) $|\lambda_k| \leqslant \left(\sum_{ij} |a_{ij}|^2\right)^{1/2}$,

b) $|\lambda_k| \leqslant n \max_{ij} |a_{ij}|$,

c) $|\lambda_k| \leqslant \max_{j} \sum_{i} |a_{ij}|$, $|\lambda_k| \leqslant \max_{i} \sum_{j} |a_{ji}|$,

d) $|\lambda_k| \leqslant \alpha_1$ pour la norme spectrale (α_1 est ici le nombre singulier maximal de la matrice A).

La dernière inégalité peut être complétée par la borne inférieure. Plus précisément, on a la

PROPOSITION 2. *Les modules des nombres caractéristiques de la matrice A sont compris entre ses nombres singuliers minimal et maximal.*

Pour le démontrer, considérons la matrice-colonne ξ satisfaisant à la condition $A\xi = \lambda_k\xi$. Soient comme toujours $\xi^* = {}^t\bar{\xi}$ et $\|\xi\|^2 = \xi^*\xi$. On supposera que $\|\xi\| = 1$. Alors

$$|\lambda_k|^2 = \|A\xi\|^2 = (A\xi)^*A\xi = \xi^*(A^*A)\xi.$$

Par ailleurs, selon la formule (9) du § 2, ch. XI, la grandeur $\xi^*(A^*A)\xi$ est comprise pour $\|\xi\| = 1$ entre les nombres caractéristiques minimal et maximal de la matrice A^*A, autrement dit entre α_1^2 et α_n^2, où α_n et α_1 sont les nombres singuliers minimal et maximal. On aboutit ainsi à la double inégalité équivalente à l'assertion nécessaire $\alpha_n^2 \leqslant |\lambda_k|^2 \leqslant \alpha_1^2$.

3. Estimations des parties réelles et imaginaires des nombres caractéristiques. En fait, le même procédé de réduction à une matrice hermitienne, utilisé lors de la démonstration de la proposition 2, peut être appliqué à la démonstration de l'assertion suivante.

PROPOSITION 3. *Posons que les matrices S et T sont définies par les égalités* $S = \frac{1}{2}(A + A^*)$ *et* $T = \frac{1}{2i}(A - A^*)$. *Dans ce cas, pour tout nom-*

bre caractéristique λ_i de la matrice A, on a

$$\mu_1 \leqslant \operatorname{Re} \lambda_i \leqslant \mu_n, \quad \nu_1 \leqslant \operatorname{Im} \lambda_i \leqslant \nu_n,$$

où μ_1, ν_1 et μ_n, ν_n sont respectivement les nombres caractéristiques minimaux et maximaux des matrices S et T.

Démontrons la première double inégalité. Soient $A\xi = \lambda_i \xi$ et $\|\xi\| = 1$. Alors $\xi^* A^* = \bar{\lambda}_i \xi^*$ et il vient

$$\xi^* S \xi = \frac{1}{2}(\xi^* A \xi + \xi^* A^* \xi) = \frac{1}{2}(\xi^* \lambda_i \xi + \xi^* \bar{\lambda}_i \xi) = (\operatorname{Re}\lambda_i)\xi^*\xi = \operatorname{Re}\lambda_i.$$

Or pour $\|\xi\| = 1$ le nombre $\xi^* S \xi$ est compris entre μ_1 et μ_n. La seconde double inégalité se démontre de la même façon.

En considérant que la matrice A est complexe, on peut profiter du théorème 3, § 2, ch. XI, selon lequel il existe une matrice unitaire U telle que la matrice $R = U^{-1} A U$ est triangulaire supérieure. Les éléments diagonaux de R sont les nombres caractéristiques λ_i de la matrice A. La norme unitaire (hermitienne) de la matrice R peut être calculée ainsi :

$$\|R\|_U^2 = \sum_{i,j} r_{ij}\bar{r}_{ij} = \sum_i |\lambda_i|^2 + \sum_{i<j} |r_{ij}|^2 \geqslant \sum_i |\lambda_i|^2.$$

Or la norme unitaire de la matrice A ne varie pas quand on multiplie A à gauche ou à droite par la matrice unitaire. Donc,

$$\|A\|_U^2 \geqslant \sum_i |\lambda_i|^2. \tag{1}$$

Ceci étant, on a l'égalité si et seulement si tous les $r_{ij} = 0$ $(i < j)$, c'est-à-dire que la matrice R est diagonale.

Recherchons quelques corollaires de l'inégalité (1).

Notons U^* la matrice ${}^t\bar{U}$. Alors pour la matrice unitaire U on a $U^* = U^{-1}$ et par suite, $R^* = (U^* A U)^* = U^* A^* U$. D'où, pour la matrice S définie dans la proposition 3,

$$U^* S U = \frac{1}{2} U^*(A + A^*) U = \frac{1}{2}(R + R^*).$$

Les éléments diagonaux de la matrice $\frac{1}{2}(R + R^*)$ sont égaux à $\operatorname{Re} \lambda_i$, $i = 1, \ldots, n$, et pour la norme unitaire de la matrice S, on obtient

$$\|S\|_U^2 = \sum_{i=1}^{n} |\operatorname{Re}\lambda_i|^2 + \frac{1}{2}\sum_{i \neq j}(s_{ij} + \bar{s}_{ij}) \geqslant \sum_{i=1}^{n} |\operatorname{Re}\lambda_i|^2.$$

De façon analogue on démontre que

$$\|T\|_U^2 \geqslant \sum_{i=1}^{n} |\mathrm{Im}\ \lambda_i|^2,$$

où la matrice T est définie dans la proposition 3.

Au § 3 du ch. XI on a vu que la c'-norme de la matrice est supérieure à sa norme euclidienne. Le même résultat est assurément vrai pour la norme unitaire :

$$\|A\|_U \leqslant n \max_{i,j} |a_{ij}| = \|A\|_{c'}.$$

En l'appliquant aux matrices S et T, on obtient

$$|\mathrm{Re}\ \lambda_i| \leqslant \left(\sum_{i=1}^{n} |\mathrm{Re}\ \lambda_i|^2\right)^{1/2} \leqslant \|S\|_U \leqslant \|S\|_{c'},$$

et finalement

$$|\mathrm{Re}\ \lambda_i| \leqslant n \max_{i,j} \left|\frac{a_{ij} + \bar{a}_{ji}}{2}\right|.$$

D'une façon analogue,

$$|\mathrm{Im}\ \lambda_i| \leqslant n \max_{i,j} \left|\frac{a_{ij} - \bar{a}_{ij}}{2}\right|.$$

On a montré que $\|A\|_U^2 = \mathrm{tr}\ A^*A$. Or

$$\mathrm{tr}\ A^*A = \sum_{i=1}^{n} \alpha_i^2,$$

où α_i sont les nombres singuliers de A. Aussi l'inégalité (1) signifie-t-elle que

$$\sum_{i=1}^{n} |\lambda_i|^2 \leqslant \sum_{i=1}^{n} \alpha_i^2,$$

et l'égalité a lieu si et seulement si A est une matrice normale.

4. Disques de localisation. On dira que la matrice A d'ordre n possède une *diagonale principale dominante* si le module de chacun de ses éléments diagonaux est strictement supérieur à la somme des modules des autres éléments de la même ligne :

$$|a_{ii}| > \sum_{k \neq i} |a_{ik}| \text{ pour tous les } i = 1, \ldots, n. \tag{2}$$

PROPOSITION 4. *Si la matrice A possède une diagonale principale dominante,* $\det A \neq 0$.

En effet, si det $A = 0$, il existe une solution non triviale du système d'équations linéaires $A\boldsymbol{\xi} = \mathbf{0}$. Soit ξ^i la composante maximale en module de cette solution. En portant $\boldsymbol{\xi}$ dans la i-ième équation du système, on obtient

$$|a_{ii}\xi^i| = \sum_{k \neq i} a_{ik}\xi^k ,$$

d'où

$$|a_{ii}||\xi^i| \leqslant \sum_{k \neq i} |a_{ik}||\xi^k| \leqslant |\xi^i| \sum_{k \neq i} |a_{ik}|,$$

ce qui est contraire à l'hypothèse.

En appliquant la proposition 4 à la matrice transposée, on voit que, pour que la matrice A soit de déterminant non nul, il suffit également que les éléments de la diagonale principale dominent les éléments des colonnes, c'est-à-dire que

$$|a_{ii}| > \sum_{k \neq i} |a_{ki}|.$$

Soit donnée une condition quelconque imposée à la matrice A, sous laquelle det $A \neq 0$. On peut l'appliquer à la matrice $A - \lambda E$ et obtenir un ensemble de points $\mathscr{S}$ du plan complexe qui ne contient aucune racine du polynôme caractéristique de A. On montrera par là que toutes les racines se trouvent dans l'ensemble complémentaire de $\mathscr{S}$. Ces raisonnements sont utilisés pour démontrer une série de théorèmes de localisation. Profitons-en, ainsi que de la proposition 4.

PROPOSITION 5. *Tous les nombres caractéristiques de la matrice A appartiennent à la réunion des disques*

$$|\lambda - a_{ii}| \leqslant \sum_{k \neq i} |a_{ik}|, \quad i = 1, \ldots, n. \tag{3}$$

En effet, si le nombre λ n'appartient pas à la réunion de ces disques, chacune des conditions (3) doit être perturbée et par suite, la matrice $A - \lambda E$ doit posséder une diagonale principale dominante.

On appellera les disques (3) *disques de localisation* de la matrice A. Si $a_{ik} = 0$ pour $k \neq i$, le i-ième disque se réduit en un point. Pour la matrice unité par exemple, la réunion des disques contient un seul point.

Pour préciser le sens des deux dernières phrases, introduisons un facteur devant les éléments non diagonaux de la matrice A et étudions la fonction matricielle $F(t)$ à éléments $f_{ii}(t) = a_{ii}$ et $f_{ik}(t) = ta_{ik}$ si $i \neq k$, définie pour $t \in [0, 1]$. Si $t \to 0$, les rayons des disques de la matrice $F(t)$ tendent vers zéro, tandis que les centres demeurent immobiles.

Utilisons, sans le démontrer, le théorème qui affirme que les racines

d'un polynôme sont des fonctions continues de ses coefficients (voir Ostrowski [30]). Les coefficients du polynôme caractéristique de la matrice $F(t)$ sont des polynômes de ses éléments et par suite, des fonctions continues de t. Pour cette raison, les nombres caractéristiques de $F(t)$ sont aussi des fonctions continues de t. Lorsque t varie de 1 à 0, chacune des racines décrit sur le plan complexe un arc continu. Plus précisément, pour chaque nombre caractéristique λ_j de la matrice A il existe un arc continu d'origine en λ_j et d'extrémité en un point a_{ii}. Ceci étant, chaque point a_{ii} est l'extrémité d'un arc au moins.

Pour nous représenter les possibilités éventuelles, considérons l'exemple suivant. Soit

$$A = \begin{Vmatrix} -1 & 1 \\ -4 & -5 \end{Vmatrix}, \quad F(t) = \begin{Vmatrix} -1 & t \\ -4t & -5 \end{Vmatrix}.$$

Les nombres caractéristiques de $F(t)$ sont $\lambda_1 = -3 + 2\sqrt{1 - t^2}$ et $\lambda_2 = -3 - 2\sqrt{1 - t^2}$. Lorsque t diminue de 1 à 0, ils varient respectivement de -3 à -1 et à $\sqrt{-5}$. Si $t > 4/5$, aucun des nombres caractéristiques ne tombe dans le petit disque $|\lambda + 1| \leqslant t$. Pour $t < 4/5$, l'un d'eux se trouve dans le petit disque et l'autre, dans le grand $|\lambda + 5| \leqslant 4t$, et les deux disques sont disjoints (fig. 53).

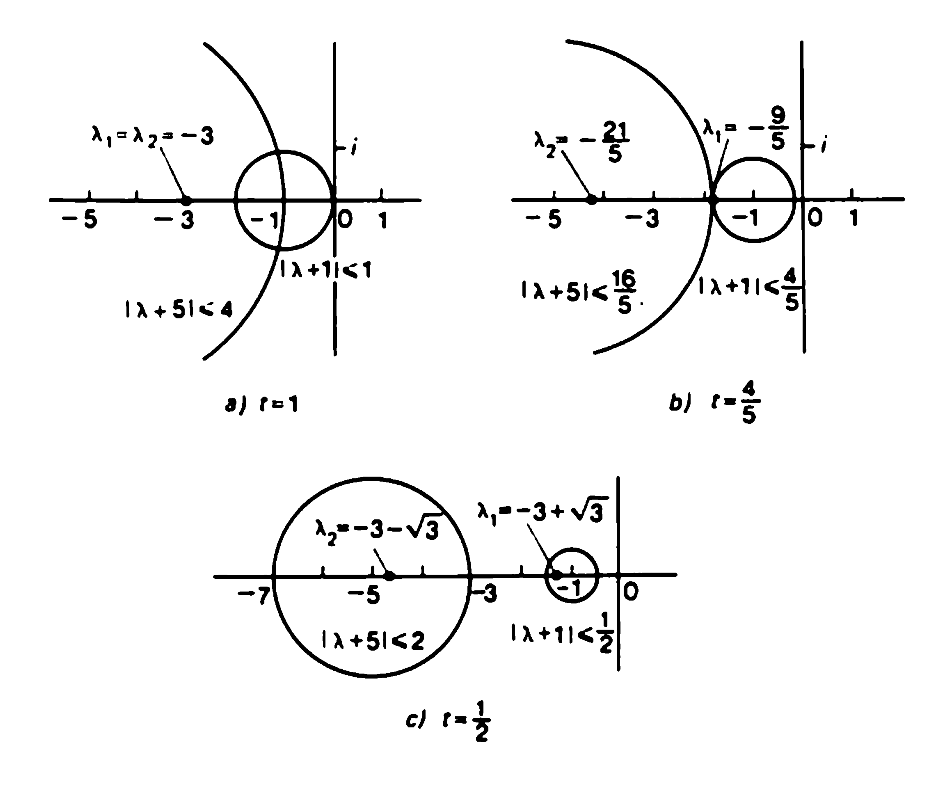


Fig. 53.

Les arcs auraient pu ne pas avoir de points communs, avoir une même extrémité et les origines différentes, etc.

Si l'un des disques de localisation de la matrice A ne coupe pas les autres, il en est de même du disque correspondant de la matrice $F(t)$ pour tout $t \in [0, 1]$. Cela signifie qu'un arc d'extrémité au centre de ce cercle y est contenu tout entier. En effet, si les disques sont disjoints, le nombre caractéristique ne peut parvenir à aucun autre disque par un mouvement continu sans passer par les points n'appartenant pas à la réunion des disques. Vu que pour tout t il doit se trouver dans un disque au moins, il restera dans le disque considéré. Ainsi donc, si l'un des disques de localisation de la matrice A ne coupe pas les autres, il contient au moins un nombre caractéristique. Il se peut qu'il en contient plusieurs, ce que montre l'exemple suivant. La matrice

$$\begin{Vmatrix} 5 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{Vmatrix}$$

possède deux disques de localisation disjoints $|\lambda - 1| \leqslant 2$ et $|\lambda - 5| \leqslant 1$. Dans le premier d'entre eux, se trouvent deux nombres caractéristiques -1 et 3.

Il est toutefois facile de remarquer que le premier disque est engendré par la deuxième et la troisième ligne de la matrice et par suite, on est en droit de le compter deux fois.

Proposition 6. *Compte tenu de la multiplicité des disques de localisation et des nombres caractéristiques, l'ensemble de r disques de localisation ne se coupant pas avec les autres $n - r$ disques comporte exactement r nombres caractéristiques.*

Cette proposition se démontre à l'aide des considérations mentionnées plus haut et on laisse au soin du lecteur d'effectuer la démonstration s'il en a l'intention.

Corollaire. *Si un disque non multiple de la matrice réelle ne se coupe pas avec les autres, il contient une racine réelle.*

En effet, dans le cas contraire, la racine complexe conjuguée aurait dû s'y trouver, vu que le centre du disque est situé sur l'axe réel.

5. Remarques et corollaires. On peut obtenir un deuxième ensemble de disques de localisation en utilisant les colonnes au lieu des lignes : tous les nombres caractéristiques de la matrice se trouvent dans la réunion des disques

$$|\lambda - a_{ii}| \leqslant \sum_{k \neq i} |a_{ki}|.$$

On peut ainsi préciser la disposition des racines.

Un autre procédé d'amélioration des estimations considérées consiste dans l'application de ces dernières à la matrice $S^{-1}AS$. Si S est une matrice diagonale, chaque ligne de A devient multipliée par un certain nombre, tandis que la colonne de même numéro, par le nombre inverse. Les centres des disques restent donc immobiles, tandis que les rayons changent de valeur. On peut essayer de choisir S de manière qu'ils diminuent.

Les autres domaines de localisation, différant des disques, peuvent être obtenus à l'aide de la proposition 9 du § 3. Soit $B = p(A)$ un polynôme de A. Il vient alors

$$|p(\lambda) - b_{ii}| \leqslant \sum_{k \neq i} |b_{ik}|. \tag{4}$$

Les domaines de ce type peuvent fournir une localisation très précise, mais son utilisation présente des difficultés. Dans le cas limite où p est un polynôme caractéristique, (4) se réduit à la condition $p(\lambda) = 0$.

Les disques de localisation permettent d'estimer les parties réelles et les modules des nombres caractéristiques. Par exemple, pour tout nombre caractéristique λ,

$$\begin{aligned} \operatorname{Re} \lambda &\leqslant \max \left[\operatorname{Re} a_{ii} + \sum_{k \neq i} |a_{ik}|\right], \\ \operatorname{Re} \lambda &\geqslant \min \left[\operatorname{Re} a_{ii} - \sum_{k \neq i} |a_{ik}|\right], \end{aligned} \tag{5}$$

car les points extrêmes gauche et droit dans le i-ième disque sont les points

$$a_{ii} - \sum_{k \neq i} |a_{ik}| \quad \text{et} \quad a_{ii} + \sum_{k \neq i} |a_{ik}|.$$

Vu que $|\lambda - a_{ii}| \geqslant |\lambda| - |a_{ii}|$, il existe un i tel que

$$|\lambda| \leqslant |a_{ii}| + \sum_{i \neq k} |a_{ik}| = \sum_{k=1}^{n} |a_{ik}|,$$

et l'on aboutit, une fois encore, à la majoration c) de la p. 365. D'une façon analogue, en utilisant les colonnes et non pas les lignes, on obtient la majoration

$$|\lambda| \leqslant \sum_{k=1}^{n} |a_{ki}|.$$

Soit u la somme maximale des modules d'éléments suivant les lignes, et v la somme maximale des modules d'éléments suivant les colonnes. On peut alors généraliser les inégalités obtenues plus haut et écrire

$$|\lambda| \leqslant \min(u, v).$$

Considérons la matrice A pour laquelle les conditions (2) ne sont pas satisfaites pour toutes les valeurs de i mais seulement pour r d'entre elles. Il découle de la proposition 4 qu'il existe dans A une sous-matrice d'ordre r avec la diagonale principale dominante, et donc, un mineur d'ordre r différent de zéro, d'où Rg $A \geqslant r$.

Supposons maintenant que pour un nombre caractéristique λ_0 de la matrice A, le rang de la matrice $A - \lambda_0 E$ vaut r. Dans ce cas, aucun mineur d'ordre $r + 1$ ne peut posséder une diagonale principale dominante et, par suite, la condition (2) est violée pour $n - r$ valeurs de l'indice i. Cela signifie que le nombre λ_0 se trouve dans les $n - r$ disques. D'où la

PROPOSITION 7. *Si à la valeur propre λ_0 de la transformation* **A** *correspondent $n - r$ vecteurs propres linéairement indépendants, λ_0 appartient à $n - r$ disques de localisation de la matrice A associée à la transformation* **A**.

Les formules (5) entraînent la condition suffisante suivant laquelle la forme quadratique est définie positive :

PROPOSITION 8. *Si la matrice associée à la forme quadratique possède une diagonale principale dominante dont les éléments sont strictement positifs, la forme quadratique est définie positive.*

En effet, selon (5), tous les nombres caractéristiques de la matrice sont strictement positifs.

Tout ce qui vient d'être dit ici suffit pour se représenter les applications possibles des disques de localisation.

Les théorèmes de localisation mentionnés dans ce paragraphe se démontrent facilement et leur application est aisée. Certains résultats importants de cette théorie ne possèdent pas lesdites propriétés et ne peuvent être exposés ici. Ceci se rapporte en particulier aux critères de Raousse-Gourvitz et de Liapounov. Ils fournissent les conditions nécessaires et suffisantes pour que tous les nombres caractéristiques de la matrice présentent des parties réelles négatives. Le lecteur peut s'initier à ces questions en s'adressant à des ouvrages spéciaux traitant de la théorie des matrices, par exemple au livre de Gantmacher [12].

CHAPITRE XIII

INTRODUCTION AUX MÉTHODES NUMÉRIQUES

§ 1. Introduction

1. Objet du chapitre. Aujourd'hui, un ingénieur ou un chercheur s'adresse tout naturellement à un ordinateur pour résoudre un système, quelque peu important, d'équations linéaires. Il existe pour cela des programmes efficaces qui fonctionnent parfaitement. Aussi pour la plupart des lecteurs de ce livre, la résolution pratique d'un système d'équations linéaires se réduit-elle à la recherche d'un programme ou d'un procédé standard. Toutefois, pour bien poser le problème, choisir le programme et interpréter les résultats, il est très important de connaître les algorithmes utilisés, ainsi que les difficultés principales auxquelles on se heurte au cours de la résolution des systèmes linéaires, et les moyens permettant de les surmonter. C'est de quoi on parlera dans le présent chapitre.

Il va de soi que le logiciel de l'ordinateur se perfectionne constamment. La mise au point de nouveaux algorithmes et programmes permettant de résoudre les problèmes d'algèbre linéaire demeure un sujet d'actualité. Mais il faut avoir en vue que la composition d'un programme plus efficace que celui qui existe déjà est une entreprise laborieuse, exigeant non seulement des connaissances profondes mais aussi une expérience pratique. C'est pourquoi, pour tous ceux qui se préparent à cette activité, le présent chapitre n'est qu'une entrée en matière. Sous ce rapport, remarquons que l'exposé suffisamment exhaustif des résultats de tout paragraphe de ce chapitre peut remplir tout un volume.

On étudiera en grands traits dans le premier paragraphe les principales difficultés auxquelles on se heurte dans la résolution numérique des problèmes d'algèbre linéaire. En fait, on passe dans le chapitre XIII au domaine des mathématiques appliquées dont les difficultés sont spécifiques à la branche.

2. Erreurs d'arrondi. Une des plus manifestes particularités des mathématiques appliquées est la prise en considération des erreurs d'arrondi. Ceci est lié au fait que le registre de l'ordinateur ne peut contenir qu'un nombre fini de caractères. Pour fixer les idées, on parlera de chiffres décimaux, bien que les ordinateurs utilisent plus souvent le système binaire ; on se sert aussi d'autres systèmes de numération qui ont pour base 8 ou 16. Les

nombres qui ne peuvent être représentés par une suite décimale finie, et de plus suffisamment courte, doivent être arrondis et, par suite, ne peuvent présenter à l'ordinateur leurs valeurs exactes. Le perfectionnement des moyens techniques peut augmenter la précision mais non pas lever le problème.

Il ne faut pas penser que l'ordinateur seul est responsable de ces défauts dans la représentation des nombres. Puisqu'il est impossible d'écrire un nombre infini de chiffres, toute suite décimale illimitée représentant un nombre réel doit être tronquée. On peut évidemment (et on le fait souvent) représenter les nombres rencontrés par des symboles tels que $\sqrt{2}$ et les manipuler ensuite en obtenant des résultats exacts de type $\pi\sqrt{2} - e^2$. Mais on ne peut utiliser ce résultat à autre chose qu'à la comparaison avec les réponses citées dans un recueil de problèmes. Pour qu'on puisse l'utiliser, on doit le calculer ou estimer et, par suite, arrondir. D'autre part, on peut se limiter aux nombres rationnels et les manipuler comme les fractions ordinaires. Mais s'il s'agit d'effectuer un nombre assez important d'opérations arithmétiques, on aboutit à des fractions dont les numérateurs et les dénominateurs sont très grands, qu'en fin de compte on sera obligé d'arrondir. Ainsi donc, dans les calculs numériques, les erreurs d'arrondi sont aussi inévitables que le sont en physique les erreurs de mesure.

Notons que les possibilités d'augmenter la précision sont en principe très grandes. En utilisant les programmes spéciaux, on peut représenter les nombres et effectuer sur eux les opérations arithmétiques avec une précision de loin supérieure à la précision ordinaire (voir, par exemple, Dreyfus, Gangloff [6]). Toutefois, même deux cents chiffres décimaux ne constituent qu'une représentation approximative des nombres. En outre, la mise en œuvre de tels procédés augmente le temps de calcul jusqu'aux grandeurs inadmissible dans un travail réel. Aussi pour les problèmes d'algèbre linéaire une telle approche n'a-t-elle pas d'intérêt pratique.

Considérons deux procédés principaux de représentation approchée des nombres.

1) *Représentation en virgule fixe*. Notons $F_{r,t}$ l'ensemble des nombres positifs et négatifs écrits avec r chiffres décimaux dont t chiffres après la virgule représentent la partie fractionnaire. L'ensemble $F_{r,t}$ est fini, bien qu'il contienne un nombre élevé d'éléments si r est assez grand. Le nombre réel a peut être représenté approximativement par un nombre de $F_{r,t}$ si la partie entière de a est inférieure à 10^{r-t}, autrement dit, peut être écrite par $r - t$ chiffres décimaux.

Dans ce cas, ou $a \in F_{r,t}$, ou il existe dans $F_{r,t}$ des nombres a' et a'', qui sont respectivement le plus grand minorant et le plus petit majorant du nombre a. Si $|a - a'| \leqslant |a - a''|$, on dit que a' est la représentation approchée de a par les nombres de $F_{r,t}$, dans le cas contraire, c'est a''.

Le nombre a se trouve dans l'intervalle $]a', a''[$ et, par suite, sa distance à l'une de ces extrémités est au plus égale à $\frac{1}{2}(a'' - a')$. Il est évident que $a'' - a' = 10^{-t}$, d'où la proposition suivante.

PROPOSITION 1. *Si le nombre a est approché par un nombre $\bar{a}$ de $F_{r,t}$, le module de l'erreur absolue $|a - \bar{a}|$ ne dépasse pas* $\frac{1}{2} 10^{-t}$.

2) *Représentation en virgule flottante.* Soit μ un nombre de l'intervalle $[0,1, 1[$ appartenant à l'ensemble $F_{d,d}$, et soit k un nombre entier de module inférieur à p. On notera $E_{d,p}$ l'ensemble des nombres de la forme $\pm\mu \cdot 10^k$. Le zéro est aussi inclus dans $E_{d,p}$. Le nombre μ est appelé *mantisse* et k *ordre* d'un nombre de $E_{d,p}$.

Si le nombre a est en module inférieur à 10^p et supérieur à 10^{-p}, il représenté par un nombre $\bar{a}$ de $E_{d,p}$. A cet effet, on l'écrit sous forme de $\pm m \cdot 10^k$, où $m \in [0,1, 1[$ et k un nombre entier. Ensuite, on représente m de façon approchée par un nombre μ de $F_{d,d}$. Dans ce cas, l'erreur absolue $a - \bar{a}$ ne dépasse par en module $\frac{1}{2} 10^{k-d}$. Pour le module de l'erreur relative on a

$$\frac{10^{k-d}}{2|a|} = \frac{10^{k-d}}{2m\,10^k} \leqslant \frac{1}{2} 10^{1-d},$$

vu que $m^{-1} \leqslant 10$. Ainsi donc, on a la proposition suivante.

PROPOSITION 2. *Si le nombre a peut être approché par un nombre de $E_{d,p}$, l'erreur relative de cette approximation ne dépasse pas en module* $\frac{1}{2} 10^{1-d}$, *où d est le nombre de chiffres de la mantisse.*

Chacune de ces représentations a ses avantages. La représentation des nombres en virgule fixe rend les opérations arithmétiques plus rapides, mais l'ensemble des nombres représentés sous cette forme est assez restreint. La représentation des nombres en virgule flottante est plus étendue, grâce à la possibilité de représenter sous cette forme les nombres qui diffèrent fortement en grandeur. Pour plus de détails, voir par exemple le livre de Voïévodine [40].

Il faut se représenter de façon absolument nette que les opérations arithmétiques ordinaires ne sont pas des opérations sur les ensembles $F_{r,t}$ ou $E_{d,p}$, vu que les résultats de ces opérations sur les éléments de l'un quelconque de ces ensembles n'appartiennent généralement pas à cet ensemble.

Aussi définit-on en fait de nouvelles opérations qui approchent les opérations arithmétiques ordinaires. On les notera $\underset{F}{\times}$, $\underset{F}{+}$, $\underset{E}{\times}$, etc. Toutes ces opérations se définissent de façon semblable. Définissons par exemple $\underset{E}{+}$.

Soient a et b des nombres de $E_{d,\,p}$. Considérons leur somme ordinaire « exacte » $c = a + b$. Si $|c| \leqslant 10^{-p}$, on pose $a \underset{E}{+} b = 0$. Dans ce cas, on dit que le résultat de l'addition est un *zéro de machine*. Si $|c| \geqslant 10^{p}$, le résultat de l'opération n'est pas défini. On dit alors qu'il y a « *dépassement* ». Ainsi, l'opération $\underset{E}{+}$ n'est pas toujours définie. (En algèbre général, une opération de ce type est dite partielle, mais on n'introduira pas ce terme.) L'éventualité d'un dépassement doit être toujours prise en compte.

Ensuite, si $10^{-p} < |c| < 10^{p}$, il existe dans $E_{d,\,p}$ un nombre $\bar{c}$ constituant une approximation de c, et l'on posera $a \underset{E}{+} b = \bar{c}$. Autrement dit, la somme exacte est arrondie et est représentée en virgule flottante. Le nombre obtenu est considéré comme valeur approchée de la somme. La différence $c - \bar{c}$ est l'erreur d'arrondi apparaissant avec l'addition.

La précision de la représentation des nombres en virgule flottante peut être caractérisée par le nombre appelé « *ε de machine* ». On le définit comme le plus petit nombre tel que $1 \underset{E}{+} \varepsilon > 1$. Voyons à titre d'exemple les opérations dans l'ensemble $E_{4,\,1}$. En ajoutant à $0{,}1000 \cdot 10^{1}$ le nombre $0{,}5000 \cdot 10^{-4}$, on obtient de nouveau $0{,}1000 \cdot 10^{1}$, tandis qu'en ajoutant $0{,}5001 \cdot 10^{-4}$, on obtient déjà $0{,}1001 \cdot 10^{1}$.

Il faut remarquer que les opérations arithmétiques approchées $\underset{E}{+}$, $\underset{E}{\times}$, etc. possèdent des propriétés absolument différentes de celles des opérations exactes. Par exemple, elles ne sont pas associatives, la loi de distributivité ne joue pas, il existe des diviseurs de zéro, c'est-à-dire que le produit de facteurs non nuls peut devenir nul (on appelle cette situation « apparition de zéro de machine en multiplication »).

Avec l'exécution d'un grand nombre d'opérations arithmétiques, les erreurs d'arrondi s'accumulent : le résultat peut différer considérablement de celui qu'on obtient par des opérations exactes. L'estimation d'erreurs engendrées (*l'analyse directe d'erreurs d'arrondi*) est une affaire compliquée et laborieuse, vu la complexité des propriétés d'opérations arithmétiques approchées. Dans certains cas, l'analyse directe d'erreurs d'arrondi peut être remplacée par les procédés utilisant le principe appelé *analyse inverse d'erreurs d'arrondi*. Selon ce principe, la solution approchée d'un problème quelconque, par exemple, d'un système d'équations linéaires, est la solution exacte d'un certain problème approché, à savoir, dans l'exemple donné, d'un système d'équations linéaires dont les coefficients sont quelque peu modifiés. Au lieu d'estimer la différence entre la solution approchée et la solution exacte, on peut estimer la différence entre les coefficients du système initial et du système modifié. L'analyse inverse d'erreurs d'arrondi s'effectue de façon beaucoup plus simple que l'analyse directe.

Donnons un exemple type de problème où l'on peut se limiter à l'analyse inverse d'erreurs. Supposons que les coefficients du système d'équations linéaires ont été obtenus par des mesures physiques et par suite, nous sont connus avec une certaine erreur. Si par analyse inverse on montre que l'influence des erreurs d'arrondi est équivalente à une altération des coefficients, inférieure aux erreurs de mesure, on peut considérer que le calcul est suffisamment exact et ne pas prêter attention à la différence entre les solutions calculée et exacte du système, lui-même déterminé de façon insuffisamment précise.

3. Influence de l'imprécision d'une information initiale. Si un problème pratique se réduit à un système d'équations linéaires, les coefficients et les termes constants du système sont en général connus avec quelques erreurs. En outre, comme il a été noté plus haut, la représentation des données initiales et les calculs présentent des erreurs d'arrondi dont l'influence équivaut à une certaine altération des coefficients et des termes constants.

On ne peut éviter les altérations mentionnées, mais on est en mesure d'estimer l'erreur obtenue et ensuite de choisir une telle méthode de résolution du système qui n'augmente pas l'incertitude du résultat, déjà en germe dans le système même.

La position adoptée sera la suivante : soit un système d'équations linéaires $Ax = b$ qu'on appellera *système initial* ou *non perturbé*. On considère encore un système, dit *perturbé*, dont on sait que les coefficients et les termes constants se trouvent dans les intervalles donnés

$$]a_{ij} - \Delta a_{ij}, a_{ij} + \Delta a_{ij}[, \quad]b_i - \Delta b_i, b_i + \Delta b_i[,$$

où a_{ij} et b_i sont respectivement les coefficients et les termes constants du système initial.

On appellera *erreur* de la solution la différence entre les solutions des systèmes initial et perturbé. (On admet que chacun des systèmes a une seule solution.) Selon cette définition, l'erreur est une matrice-colonne, de sorte qu'elle peut être estimée en une certaine norme de l'espace arithmétique. Dans les exemples de ce paragraphe, on étudie exclusivement la norme euclidienne.

Considérons le plus simple des exemples. Soit un système de deux équations linéaires à deux inconnues. Le système non perturbé est de la forme

$$\begin{aligned} ax + by &= c, \\ a_1 x + b_1 y &= c_1. \end{aligned} \qquad (1)$$

Admettons que les coefficients affectant les variables sont exactes et ne sont perturbés que les termes constants, c'est-à-dire que $\Delta a = \Delta b = \Delta a_1 = \Delta b_1 = 0$.

Géométriquement, chaque équation du système non perturbé est repré-

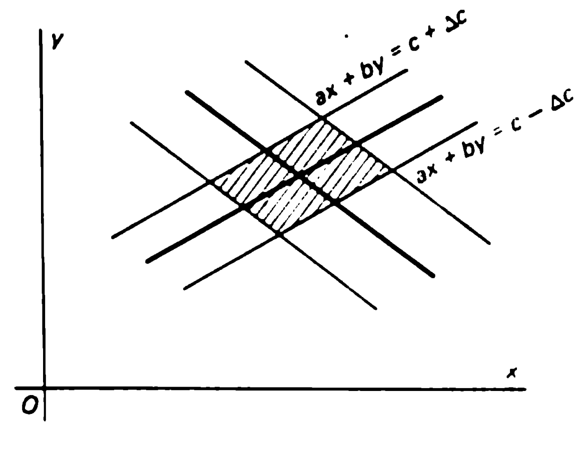


Fig. 54.

sentée par une droite du plan. L'équation correspondante du système perturbé se représente alors par une droite qui lui est parallèle et qui est située à l'intérieur d'une bande (fig. 54). Pour la première équation, la bande est limitée par les droites

$$ax + by = c - \Delta c \quad \text{et} \quad ax + by = c + \Delta c.$$

Ainsi, les solutions du système perturbé sont situées à l'intérieur du parallélogramme formé par l'intersection des bandes.

L'erreur de la solution est représentée par le vecteur $(\Delta x, \Delta y)$ dont la longueur peut atteindre la moitié de celle de la plus grande diagonale du parallélogramme. Par conséquent, l'influence due aux perturbations des termes constants du système est d'autant plus grande pour des bandes de même largeur que l'angle des droites (1) est plus petit.

Dans le cas général où les coefficients du système perturbé diffèrent aussi de ceux du système initial, les droites représentant les équations subissent non seulement des translations mais encore des rotations. Le parallélogramme se remplace par une figure plus compliquée, mais le résultat général demeure le même : plus l'angle des droites (1) est petit, moins bien le système est conditionné, c'est-à-dire qu'avec la même perturbation des coefficients, l'erreur de la solution peut devenir plus grande.

Notre objectif est de définir exactement et, si possible, de mesurer quantitativement la propriété d'un système d'équations linéaires d'être plus ou moins bien conditionné.

On voit bien sur l'exemple considéré que si on remplace le système donné par un système équivalent, son conditionnement varie car dans ce cas le couple de droites (1) se remplace par un autre couple de droites ayant le même point d'intersection. On pourrait croire que le système le mieux conditionné est celui dont les droites représentatives sont perpendiculaires. Or une étude plus détaillée, tenant compte de la largeur des bandes, montre

qu'il n'en est pas toujours ainsi. Considérons par exemple le système

$$10^3 x = 1,$$
$$10^{-3} y = -10^{-3}.$$

Une variation relativement petite des termes constants $\Delta c_1 = 0$, $\Delta c_2 = 10^{-3}$, pour laquelle

$$\frac{\|\Delta c\|}{\|c\|} = \frac{\sqrt{(\Delta c_1)^2 + (\Delta c_2)^2}}{\sqrt{c_1^2 + c_2^2}} \approx 10^{-3},$$

aboutit à l'erreur de solution égale à (0, 1) et à l'erreur relative

$$\frac{\|\Delta x\|}{\|x\|} = \frac{\sqrt{(\Delta x)^2 + (\Delta y)^2}}{\sqrt{x^2 + y^2}} \approx 1.$$

On remarque facilement que le mauvais conditionnement de ce système est dû au fait que ses coefficients et termes constants sont de grandeur différente. La situation peut être améliorée par multiplication de la seconde équation par 10^3.

Examinons la question dans une autre optique. Le système d'équations linéaires peut être interprété d'une façon différente, comme un problème qui consiste à rechercher un antécédent du point $C'(c_1 + \Delta c_1, c_2 + \Delta c_2)$ dans la transformation affine du plan

$$x^* = ax + by,$$
$$y^* = a_1 x + b_1 y. \tag{2}$$

Supposons que la transformation affine est connue de façon exacte : $\Delta a = \Delta a_1 = \Delta b = \Delta b_1 = 0$. Quant au point C', on sait seulement qu'il se trouve à l'intérieur d'un cercle de rayon ρ et de centre $C(c_1, c_2)$. L'antécédent du cercle est l'ellipse et, par suite, la solution du système perturbé se trouve à l'intérieur d'une ellipse. Le centre de cette ellipse est l'antécédent du centre du cercle, c'est-à-dire la solution du système non perturbé.

Dans une transformation affine, les antécédents de tous les cercles ont un même rapport des demi-axes. Soient $\lambda\rho$ et $\mu\rho$ ($\lambda > \mu$) les demi-axes de l'antécédent d'un cercle de rayon ρ. Dans ce cas, la longueur δ du vecteur erreur $(\Delta x, \Delta y)$ ne dépasse pas $\lambda\rho$.

Pour estimer l'erreur relative, minorons la longueur du vecteur représentant la solution du système non perturbé. Etant donné que le point C se trouve sur le cercle de rayon $r = \sqrt{c_1^2 + c_2^2}$ et de centre à l'origine des coordonnées, son antécédent est un point de coordonnées (x_0, y_0) dont la distance à l'origine des coordonnées est au moins égale à μr. Donc, $d = \sqrt{x_0^2 + y_0^2} \geqslant \mu r$, et pour l'erreur relative de la solution on obtient

l'estimation

$$\frac{\delta}{d} \leqslant \frac{\lambda}{\mu}\frac{\rho}{r}. \tag{3}$$

ρ/r est ici l'erreur relative maximale de la matrice-colonne des termes constants.

Ainsi donc, le rapport des demi-axes de l'ellipse λ/μ peut être interprété comme le nombre caractérisant le conditionnement du système (1).

On voit ici en particulier qu'un mauvais conditionnement du système est indépendant de la valeur du déterminant : il dépend du rapport des demi-axes de l'ellipse ; quant au déterminant, il est égal au rapport des aires du cercle et de l'ellipse et n'est en aucune façon lié à la forme de l'ellipse.

4. Matrices quasi singulières. La perturbation des coefficients d'un système d'équations linéaires peut non seulement altérer quelque peu sa solution mais avoir aussi des conséquences plus sérieuses. Le système d'équations linéaires

$$\begin{aligned} x + 0{,}99y &= 1{,}01, \\ x + 1{,}01y &= 0{,}99 \end{aligned}$$

a une solution unique, mais une variation de 1 % de ses coefficients et seconds membres peut aboutir soit à un système incompatible, soit à un système admettant une infinité de solutions.

Dans l'exemple donné, le déterminant de la matrice du système est petit devant ses coefficients. Il n'est pas toutefois difficile de fournir un exemple de matrice dont le déterminant n'est pas du tout petit, mais qui s'annule après une petite perturbation des éléments de la matrice. Considérons par exemple une matrice diagonale, disons d'ordre douze, où tous les éléments diagonaux sont égaux à dix, à l'exception du dernier qui est égal à 10^{-10}. Le déterminant d'une telle matrice est 10, mais peut devenir nul si on varie un seul élément de 10^{-10}.

La raison de ce fait est la suivante. Considérons les éléments de la matrice comme des variables indépendantes. Le déterminant est une fonction linéaire de chacune de ces variables, le coefficient de la variable α_{ij} étant le cofacteur de cet élément. Quelques mineurs d'ordre $n - 1$ peuvent s'avérer grands par rapport au déterminant. Dans ce cas une petite variation d'un élément peut entraîner l'annulation du déterminant. Cette question sera traitée en détail au § 2.

Pour l'instant, il nous faut souligner l'importance de principe de ce phénomène. Si les nombres ne peuvent être définis exactement, la frontière entre les matrices singulières et régulières cesse d'être nette et il apparaît une classe de matrices *quasi singulières* dont les frontières dépendent du degré de précision adoptée dans l'étude concrète. A propos de la matrice

quasi singulière, on ne peut dire si elle est singulière ou régulière, car en variant ses éléments dans les limites de précision adoptée, on peut obtenir soit une matrice singulière, soit une matrice régulière. Aussi dans les calculs approchés faut-il faire attention aux expressions du genre « admettons que la matrice A est régulière... » dont abonde le cours d'algèbre linéaire.

Si dans un système d'équations linéaires la matrice s'est avérée quasi singulière, il vaut mieux revenir au problème qui a conduit au système d'équations linéaires considéré et obtenir une information complémentaire sur ce système. Il peut arriver par exemple que d'après la situation envisagée dans le problème la solution doit être unique. Une certaine approche de ce qu'on doit considérer dans ce cas comme solution du problème sera discuté au chapitre XIV, mais ce n'est qu'une des possibilités. En réalité, il faut se représenter de façon nette qu'une indétermination quantitative dans l'information initiale exerce dans ce cas une influence qualitative sur la solution, et il serait plus correcte de concentrer les efforts non pas sur la résolution du système mais sur la position plus rigoureuse du problème.

Il n'est pas difficile de remarquer que l'apparition de la classe des matrices quasi singulières est plus en rapport avec la nature des calculs approchés qu'avec les propriétés du déterminant, envisagé comme fonction de matrice. Vu que cette fonction est continue, l'ensemble des matrices régulières est ouvert (voir Koudriavtsev [21], t. I, p. 329). Cela signifie que pour toute matrice régulière A_0 il existe un voisinage $\| A - A_0 \| < \varepsilon$ composé de matrices régulières. Or, on ne peut considérer des voisinages aussi petits que l'on veut, car le rayon du voisinage ne peut être inférieur à un certain nombre ε_0 défini par la précision des calculs ou par la valeur de l'éventuelle perturbation de la matrice. La matrice quasi singulière est la matrice dont le voisinage de rayon ε_0 contient une matrice singulière.

Ces considérations de nature assez générale montrent que le long de la frontière de tout ensemble ouvert $\mathscr{A}$ apparaît une « bande d'indétermination » de largeur ε_0, composée de ε_0-voisinages des points frontières de l'ensemble $\mathscr{A}$. Au-delà de cette bande, l'appartenance à l'ensemble peut être vérifiée numériquement. Au contraire, l'appartenance à la frontière de $\mathscr{A}$ ne peut pas être vérifiée numériquement : un tout petit écart fait passer d'un point frontière à un point de l'ensemble.

Si l'on tient compte des erreurs d'arrondi et des perturbations, on voit devenir moins nettes les frontières des ensembles tels que l'ensemble des matrices de rang fixé, celui des matrices de structure simple (c'est-à-dire des matrices dont le polynôme minimal ne possède pas de racines multiples). On voit aussi devenir moins nette la notion de système libre de vecteurs, et beaucoup d'autres choses.

On pourrait dire que l'algèbre linéaire joue par rapport à tout ceci le même rôle de modèle abstrait que la géométrie euclidienne pour le dessin technique.

5. Capacité limitée de la mémoire. La source des difficultés liées aux erreurs d'arrondi est le nombre fini de chiffres significatifs qu'on utilise pour noter un nombre. Dans la pratique de résolution des systèmes d'équations linéaires, il existe encore une source de difficultés, de nature à première vue quantitative, mais qui est en fait d'une importance de principe. Il s'agit de la capacité de la mémoire de l'ordinateur, qui limite les dimensions (c'est-à-dire le nombre d'équations et de variables) des systèmes à résoudre. Il est vrai que la capacité des mémoires externes est si grande qu'elle n'impose pas de sensibles limitations, mais l'utilisation de ces dispositifs est liée à d'autres difficultés, en premier lieu à un temps d'accès beaucoup plus grand.

Le lecteur non initié à la programmation pourrait le plus simplement se représenter la situation ainsi. Vous résolvez un système d'équations linéaires au tableau noir. Ce dernier est couvert de quadrillage dont chaque case peut contenir un nombre. Une fois le tableau couvert de nombres, on ne peut y continuer d'écrire qu'en effaçant quelque chose. Il y a encore un cahier, mais il faut du temps pour aller le chercher, d'ailleurs on ne peut recopier du tableau sur le cahier ou inversement qu'un nombre entier de pages. Il est évident que résoudre dans ces conditions un système dont la matrice ne se case pas sur le tableau ou le remplit presque en entier exige une certaine ingéniosité et un temps supplémentaire.

Bien que la capacité des mémoires et la vitesse de fonctionnement des ordinateurs nouvellement créés augmentent rapidement, les besoins de la pratique s'accroissent encore plus vite. Aussi la résolution de systèmes d'équations un peu trop grands pour un ordinateur donné est-il toujours un problème d'actualité.

Il va de soi que dans la réalisation de tout algorithme, on attache beaucoup d'attention à l'économie de place dans la mémoire vive mais, une fois sa capacité fixée, il existe assurément un ordre maximal de systèmes d'équations linéaires de forme générale qu'on est en mesure de résoudre sans recourir trop souvent aux mémoires externes. La résolution des systèmes d'ordre plus grand s'avère possible sous des conditions supplémentaires qui permettent d'utiliser des méthodes de résolution spéciales.

Il existe les classes principales suivantes des matrices dont le stockage exige moins de place que celui des matrices de même ordre et de forme générale.

a) *Matrices calculables.* Supposons que les éléments de la matrice peuvent être facilement calculés d'après des données initiales quelconques occupant relativement peu de place. Dans ce cas, au lieu de mémoriser les éléments de la matrice, on peut chaque fois les calculer. Cela accroît évidemment le temps de calcul, mais souvent ce subterfuge s'avère payant.

b) *Matrices creuses.* On appelle ainsi les matrices dont la plupart des

éléments sont nuls et dont les éléments non nuls se disposent d'une façon quelconque. On peut mémoriser seulement les éléments non nuls de la matrice creuse, mais il faudra dépenser un temps supplémentaire et une partie de la mémoire pour indiquer dans quelle ligne et quelle colonne se trouve l'éléments non nul donné.

Le schéma le plus simple consistera par exemple à mémoriser un élément de la matrice avec les numéros de ses ligne et colonne. Si on le fait de la façon la plus simple mais non pas la plus économe en réservant une cellule au stockage d'un numéro, le nombre total de cellules nécessaires s'avérera trois fois plus grand que celui d'éléments non nuls. Si, de plus, la matrice n'est remplie qu'à 1/5, il n'est pas encore clair si l'économie réalisée vaut les dépenses de temps et la complication du programme. Pratiquement, une matrice de grande dimension mérite d'être considérée creuse si le nombre d'éléments non nuls est de même ordre que le nombre de lignes dans la matrice.

Les matrices creuses apparaissent dans nombre de problèmes des mathématiques appliquées. La description du mode de stockage et de traitement de ces matrices peut être trouvée dans le livre de Twearson [37].

c) Une position intermédiaire entre les deux classes de matrices décrites plus haut est occupée par les matrices qu'on peut appeler matrices de formes spéciales. Leurs éléments nuls ont des places connues d'avance, ou bien une partie de leurs éléments se calcule d'après les éléments mémorisés. Viennent à l'esprit, par exemple, les matrices diagonales, symétriques ou triangulaires.

Parmi les matrices de formes spéciales, non rencontrées encore, il faut nommer les matrices en bande. La matrice A est dite *en bande* si ses éléments sont tels que $a_{ij} = 0$ pour $|i - j| > k$, où k est un nombre fixé. Cela signifie que ne peuvent être différents de zéro que les éléments de la matrice qui sont situés à l'intérieur de la bande de largeur $2k + 1$, orientée le long de la diagonale principale.

Les matrices en bande se rencontrent dans les diverses applications. La raison en est la suivante. Supposons qu'un système d'équations décrit des liaisons existant entre les parties composantes d'un objet réel. Si le nombre des parties composantes est grand, il est naturel que toutes les parties ne sont pas liées les unes aux autres de façon directe. Si sont directement liées les seules parties dont les numéros diffèrent de k au plus, la matrice des liaisons sera une matrice en bande.

Pour les matrices de formes spéciales les plus fréquentes, on a mis au point des modifications de la plupart des algorithmes connus ainsi que des algorithmes spéciaux permettant d'utiliser ces matrices. C'est ainsi qu'il existe des algorithmes permettant de résoudre les systèmes d'équations linéaires dont les matrices sont symétriques ou en bande, et de trouver les

valeurs propres et les vecteurs propres des transformations définies par de telles matrices. Ces algorithmes fonctionnent de façon plus efficace et permettent de résoudre des problèmes de dimension plus grande grâce à la forme spéciale de ces matrices. On n'étudiera pas ces algorithmes.

Le choix de la méthode de résolution d'un système d'équations linéaires et, en général, de tout problème concernant les matrices, dépend de façon essentielle du mode de stockage des matrices. Les méthodes, telle la méthode de Gauss, conviennent pour des matrices de forme générale, pas trop grandes. Pour des matrices encombrantes dont le stockage utilise leurs propriétés particulières, les méthodes de résolution des systèmes d'équations, basées sur les transformations de la matrice du système, sont d'une application peu commode. Au cours des transformations, la matrice creuse risque de perdre sa propriété si des mesures spéciales ne sont pas prises. De même, si les éléments d'une matrice se calculent d'après les formules simples, les éléments de la matrice transformée ne sont plus, en général, munis de cette propriété. Aussi pour résoudre les systèmes d'équations linéaires dont les matrices sont assez grandes recourt-on à des méthodes spéciales, par exemple aux méthodes itératives, étudiées au § 4.

§ 2. Conditionnement

1. Majoration de la perturbation. Soit donné le système d'équations linéaires initial

$$Ax = b, \tag{1}$$

où A est une matrice carrée d'ordre n et $\det A \neq 0$. Considérons le système perturbé

$$(A + \delta A)y = b + \delta b. \tag{2}$$

Il n'est d'abord pas clair si le système (2) aura une solution unique comme le système (1). On en imposera plus bas à δA une condition suffisante. Notre objectif immédiat est d'estimer, sous cette condition, la norme de la différence entre les solutions des deux systèmes.

Dans le présent paragraphe, on entend par norme matricielle la norme qui possède la propriété annulaire et qui est compatible avec la norme sur l'espace des matrices-colonnes.

PROPOSITION 1. *Supposons que pour une norme matricielle la matrice carrée B satisfait à la condition* $\|B\| \leqslant \rho < 1$. *Il existe alors une matrice* $(E + B)^{-1}$ *telle que* $\|(E + B)^{-1}\| \leqslant (1 - \rho)^{-1}$.

DÉMONSTRATION *). La majoration du rayon spectral de la matrice (pro-

---

*) On utilise dans la démonstration les résultats du ch. XII. Le lecteur qui n'a pas étudié ce chapitre peut adopter la proposition 1 sans démonstration.

position 1, § 4, ch. XII) entraîne que toutes les valeurs propres de la matrice B se trouvent à l'intérieur du disque $|\lambda| \leqslant \rho$, c'est-à-dire à l'intérieur du disque de convergence de la série entière en λ de la fonction $(1 + \lambda)^{-1}$. Cela garantit l'existence de la matrice $(E + B)^{-1}$ égale à la somme de la série $E - B + B^2 - \ldots$

Pour la somme partielle de cette série on a, selon la relation (12) du § 3, ch. XI, la majoration

$$\|S_k\| \leqslant \sum_0^k \|B^k\| \leqslant \sum_0^k \|B\|^k < \frac{1}{1 - \rho}.$$

D'où, en passant à la limite pour $k \to \infty$, on obtient l'inégalité exigée.

Majorons maintenant la norme de la perturbation de la solution, c'est-à-dire $\|\delta x\| = \|y - x\|$, où y est la solution du système (2) et x, du système (1). Retranchons pour cela (1) de (2). Il vient

$$(A + \delta A)(x + \delta x) - Ax = \delta b,$$

ou

$$(A + \delta A)\delta x + \delta A x = \delta b. \qquad (3)$$

Ecrivons $A + \delta A$ sous forme de $A(E + A^{-1}\delta A)$ et supposons que

$$\|A^{-1}\| \cdot \|\delta A\| = \rho < 1. \qquad (4)$$

Alors $\|A^{-1}\delta A\| \leqslant \rho$ et, selon la proposition 1, la matrice $E + A^{-1}\delta A$ possède une inverse. Donc,

$$(A + \delta A)^{-1} = (E + A^{-1}\delta A)^{-1}A^{-1},$$

et il découle de (3) que

$$\delta x = (E + A^{-1}\delta A)^{-1}A^{-1}\delta b - (E + A^{-1}\delta A)^{-1}A^{-1}\delta A x,$$

d'où

$$\|\delta x\| \leqslant \frac{1}{1 - \rho}\|A^{-1}\| \cdot \|\delta b\| - \frac{1}{1 - \rho}\|A^{-1}\| \cdot \|\delta A\| \cdot \|x\|.$$

Il ressort de l'égalité (1) que $\|b\| \leqslant \|A\| \cdot \|x\|$. Renforçons l'inégalité en multipliant le premier terme du second membre par $\|A\| \cdot \|x\|/\|b\|$. Alors,

$$\|\delta x\| \leqslant \frac{c(A)\|x\|}{1 - \rho}\frac{\|\delta b\|}{\|b\|} - \frac{c(A)\|x\|}{1 - \rho}\frac{\|\delta A\|}{\|A\|},$$

où

$$c(A) = \|A^{-1}\| \cdot \|A\|. \qquad (5)$$

Divisons par $\|x\|$ et tenons compte de ce que

$$\rho = \|\delta A\| \cdot \|A^{-1}\| = c(A)\frac{\|\delta A\|}{\|A\|}.$$

On obtient la majoration définitive

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{c(A)}{1 - c(A)\dfrac{\|\delta A\|}{\|A\|}}\left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|}\right).$$

Le nombre $c(A)$ introduit par la formule (5) est appelé *nombre conditionnel* de la matrice A en norme considérée.

On peut maintenant formuler le résultat suivant.

PROPOSITION 2. *Supposons que la matrice des coefficients et la matrice-colonne des seconds membres du système* (1) *ont subi les perturbations* δA *et* δb, *et que de plus* $\|A^{-1}\| \cdot \|\delta A\| < 1$ *en une norme matricielle. Le système perturbé possède alors une solution unique, et la perturbation relative* $\|\delta x\|/\|x\|$ *de la solution du système* (1) *est majorée au moyen des perturbations relatives* $\alpha = \|\delta A\|/\|A\|$ *et* $\beta = \|\delta b\|/\|b\|$ *de la matrice du système et de la matrice-colonne des seconds membres par la formule*

$$\frac{\|\delta x\|}{\|x\|} \leqslant \frac{c(A)(\alpha + \beta)}{1 - c(A)\alpha}, \tag{6}$$

où $c(A)$ *est le nombre conditionnel de la matrice* A *en norme considérée.*

Il est remarquable que le second membre de (6) ne contient que les perturbations relatives de A et b. La matrice A n'est représentée que par son nombre conditionnel, quant à la matrice-colonne b, elle n'y figure pas du tout.

2. Nombre conditionnel. Selon la proposition 2, plus le nombre conditionnel est grand, plus la perturbation relative de la solution est grande pour les mêmes perturbations relatives des données initiales. Le nombre conditionnel est déterminé non seulement par la matrice, mais également par le choix de la norme. Pour des normes différentes, la formule (6) donnera des majorations différentes, plus ou moins précises de la perturbation relative. Dans ce point, on examinera le calcul du nombre conditionnel de la matrice et déduira quelques-unes de ses propriétés.

Commençons par les relations simples suivantes qui sont vérifiées pour toute norme. Il découle immédiatement de la définition que

$$c(A) = c(A^{-1}). \tag{7}$$

En multipliant les inégalités $\|AB\| \leqslant \|A\| \cdot \|B\|$ et $\|(AB)^{-1}\| = \|B^{-1}A^{-1}\| \leqslant \|B^{-1}\| \cdot \|A^{-1}\|$, on obtient

$$c(AB) \leqslant c(A)c(B). \qquad (8)$$

Ensuite, à partir de $A^{-1}A = E$ on a $c(A) \geqslant \|E\|$. Vu que $\|E\| \geqslant 1$ (voir (11), § 3, ch. XI), on obtient

$$c(A) \geqslant \|E\| \geqslant 1. \qquad (9)$$

Soient α_1 et α_n le plus grand et le plus petit nombre singulier de la matrice A. Selon la proposition 14 du § 3, ch. XI, on a pour la norme spectrale $\|A\| = \alpha_1$ et, suivant la formule (16) du § 3, ch. XI, $\|A^{-1}\| = \alpha_n^{-1}$. Donc, le nombre conditionnel en norme spectrale (ou, comme on dit, *nombre conditionnel spectral*) est obtenu par la formule

$$c(A) = \frac{\alpha_1}{\alpha_n}. \qquad (10)$$

Au § 1 on a obtenu à partir des considérations géométriques que le conditionnement du système de deux équations à deux inconnues est caractérisé par le rapport des demi-axes d'une ellipse (l'antécédent du cercle par transformation linéaire définie par la matrice du système). En se rappelant la signification géométrique des nombres singuliers et la formule (10), on remarque que ce rapport est justement le nombre conditionnel spectral.

Il ressort immédiatement de (10) que pour la matrice orthogonale (ainsi que pour la matrice unitaire) le nombre conditionnel spectral est 1.

De la propriété correspondante de la norme spectrale (proposition 13, § 3, ch. XI) il découle que le nombre conditionnel spectral $c(A)$ ne varie pas quand on multiplie A par la matrice orthogonale (resp. unitaire). En vertu de ce fait, en résolvant les systèmes d'équations linéaires, il est préférable de multiplier la matrice du système par des matrices orthogonales.

Considérons maintenant les inégalités entre les nombres conditionnels associés aux normes différentes. Si deux normes φ et ψ sont telles que pour toute matrice A on a $\varphi(A) \leqslant \beta\psi(A)$, les nombres conditionnels correspondants sont liés, comme on le voit facilement, par l'inégalité

$$c_\varphi \leqslant \beta^2 c_\psi.$$

La norme euclidienne de la matrice est égale à la racine carrée de la somme des carrés de ses nombres singuliers, quant à la norme spectrale, elle est égale au nombre singulier maximal. Aussi de l'inégalité

$$\alpha_1^2 \leqslant \sum_{i=1}^{n} \alpha_i^2 \leqslant n\alpha_1^2$$

s'ensuit-il que $\|A\| \leqslant \|A\|_E \leqslant \sqrt{n}\|A\|$ et, par suite,

$$c(A) \leqslant c_E(A) \leqslant nc(A). \qquad (11)$$

D'une façon analogue, à partir de

$$\max_{i,j} |a_{ij}|^2 \leqslant \sum_{i,j} |a_{ij}|^2 \leqslant n^2 \max_{i,j} |a_{ij}|^2$$

on obtient les inégalités pour la norme $\|A\|_{c'} = n \max_{i,j} |a_{ij}|$ et la norme euclidienne

$$\frac{1}{\sqrt{n}} \|A\|_{c'} \leqslant \|A\|_E \leqslant \|A\|_{c'},$$

et, à partir de ces dernières, les estimations pour les nombres conditionnels

$$\frac{1}{n} c_{c'} \leqslant c_E \leqslant c_{c'}. \tag{12}$$

Il n'existe pas pour le calcul de $c_{c'}$ de formules satisfaisantes, commodes pour des études théoriques, mais pour une matrice concrète ce nombre peut être aisément calculé si est connue la matrice inverse. Par contre, pour le nombre conditionnel spectral il existe une formule générale satisfaisante qui est toutefois difficilement calculable pour une matrice concrète de forme générale.

L'application pratique des nombres conditionnels pour estimer l'erreur au cours de la résolution d'un système concret est rendue difficile par le fait qu'en résolvant le système on ne calcule pas la matrice inverse, de sorte que le nombre conditionnel ne peut être obtenu qu'au bout d'efforts relativement grands.

3. Matrices quasi singulières. On dit qu'une matrice est quasi singulière si par une petite variation de ses éléments elle peut être transformée en une matrice singulière. Ceci étant, la classe des matrices quasi singulières est fonction de l'ordre de petitesse adopté pour les éléments. L'exemple de la p. 380 montre que la petitesse du déterminant n'est pas une condition nécessaire pour que la matrice soit quasi singulière. Cette condition n'est non plus suffisante, comme le montre l'exemple de la matrice $\frac{1}{2} E$, où E est la matrice unité, disons d'ordre 20. On montrera plus loin le lien entre la propriété de la matrice d'être quasi singulière, la norme de sa matrice inverse et le nombre conditionnel.

Soit une matrice A. Décomposons son déterminant suivant la i-ième ligne

$$\det A = (-1)^{i+j} a_{ij} M_j^i + N,$$

où N est la somme des termes qui ne contiennent pas l'élément a_{ij}. L'addition de ε_{ij} à l'élément a_{ij} pour annuler le déterminant doit vérifier l'équa-

tion

$$(-1)^{i+j}(a_{ij} + \varepsilon_{ij})M_j^i + N = 0,$$

ou

$$(-1)^{i+j}M_j^i\varepsilon_{ij} = -\det A.$$

Si $M_j^i = 0$, il est évident que $\det A$ ne varie pas avec la modification de a_{ij}. Dans le cas contraire, on obtient

$$\varepsilon_{ij} = (-1)^{i+j+1}\frac{\det A}{M_j^i} = -(b_{ji})^{-1},$$

où b_{ji} désigne un élément de la matrice inverse A^{-1}.

On s'intéressera à la perturbation minimale en module d'un élément de la matrice, qui annule le déterminant. On a pour cette perturbation

$$\varepsilon = \min_{ij} |\varepsilon_{ij}| = \left(\max_{ij} |b_{ji}|\right)^{-1}.$$

Le minimum est ici pris suivant i, j pour lesquels $M_j^i \neq 0$. Mais le maximum peut être pris suivant tous les i, j, car pour $M_j^i = 0$ on a $b_{ij} = 0$. En y portant $\max |b_{ji}| = \frac{1}{n} \|A^{-1}\|_{c^\cdot}$ qui figure dans la définition de la norme $\| * \|_{c^\cdot}$, on obtient

$$\varepsilon = n\|A^{-1}\|_{c^\cdot}^{-1}.$$

Supposons que le minimum est atteint pour la k-ième ligne et la l-ième colonne. Notons E_{kl} la matrice dont tous les éléments sont nuls, sauf l'élément $e_{kl} = 1$. Alors, $\det(A + \varepsilon E_{kl}) = 0$ et la norme de la perturbation εE_{kl} vaut

$$\|\varepsilon E_{kl}\|_{c^\cdot} = n\varepsilon = \frac{n^2}{\|A^{-1}\|_{c^\cdot}}.$$

On a étudié ici un cas particulier quand ne varie qu'un élément de la matrice. Dans le cas général, le déterminant s'annule pour des perturbations de norme inférieure. On peut toutefois énoncer la proposition suivante.

Proposition 3. *Le déterminant de la matrice A peut être annulé par adjonction d'une matrice dont la norme ne dépasse pas $n^2\|A^{-1}\|_{c^\cdot}^{-1}$.*

Remarquons que l'estimation de la perturbation relative de la matrice A contient le nombre conditionnel :

$$\frac{\|\varepsilon E_{kl}\|_{c^\cdot}}{\|A\|_{c^\cdot}} = \frac{n^2}{c_{c^\cdot}(A)}.$$

En rapport avec la proposition 3, il faut rappeler la condition (4) selon

laquelle la perturbation δA satisfaisant à la condition $\|\delta A\| \leqslant \rho \|A^{-1}\|^{-1}$, où $\rho < 1$, ne peut annuler le déterminant de A.

Ainsi donc, la perturbation nécessaire pour annuler le déterminant est égale en norme à $\alpha \|A^{-1}\|_c^{-1}$, où $1 \leqslant \alpha \leqslant n^2$, ce qui monte que le nombre $\|A^{-1}\|_c^{-1}$ peut servir à dire dans quelle mesure la matrice A est proche d'une matrice singulière.

Si $\| * \|$ est la norme sur l'espace arithmétique, la fonction

$$g(A) = \inf_{\xi \neq 0} \frac{\|A\xi\|}{\|\xi\|}$$

est égale, selon la proposition 12 du § 3, ch. XI, à zéro pour les matrices singulières et à $\|A^{-1}\|^{-1}$ (au cas de norme induite) pour A régulière. En particulier, si la norme $\| * \|$ est euclidienne, on a

$$\|A^{-1}\|^{-1} = \alpha_n,$$

où α_n est le nombre singulier minimal de la matrice A (formule (16), § 3, ch. XI) et par suite, le nombre singulier minimal peut servir de mesure de proximité de la matrice A à une matrice singulière. Cette conclusion peut s'appuyer aussi sur les raisonnements suivants. Le théorème 1m du § 1, ch. XI, montre que pour la matrice A il existe des matrices orthogonales S et P telles que $A' = SAP$ est une matrice diagonale ayant les nombres singuliers sur la diagonale principale. La matrice A' peut être rendue singulière par une perturbation de norme égale au nombre singulier minimal α_n. Soit $\det (A' + F') = 0$. Alors

$$\det (A' + F') = \det S \det P \det (A - S^{-1}F'P^{-1}) = 0,$$

et, par suite, A peut être rendue singulière par l'adjonction de la matrice $F = S^{-1}F'P^{-1}$. Or la multiplication par une matrice orthogonale ne modifie pas la norme spectrale. Donc, $\|F\| = \alpha_n$.

On voit que la matrice A peut être rendue singulière par une perturbation de norme égale à α_n. La perturbation relative correspondante est égale en norme à

$$\frac{\|F\|}{\|A\|} = \frac{1}{\|A\| \cdot \|A^{-1}\|}.$$

Il en découle la proposition suivante.

PROPOSITION 4. *La matrice A peut être rendue singulière par une perturbation relative de norme égale à* $[c(A)]^{-1}$, *où* $c(A)$ *est le nombre conditionnel spectral de* A.

Etudions en détail l'influence de la perturbation sur la solution du système d'équations linéaires à matrice quasi singulière.

Une matrice carrée A peut être décomposée suivant le théorème 1m du

§ 1, ch. XI, en produit $A = PDQ$, dans lequel P et Q sont des matrices orthogonales et D une matrice diagonale avec nombres singuliers de la matrice A sur la diagonale. Considérons le système d'équations perturbé

$$(A + \delta A)(x + \delta x) = b + \delta b,$$

et portons-y la décomposition indiquée de A. Alors

$$(PDQ + P'P\delta A'QQ)(x + \delta x) = b + \delta b,$$

d'où

$$(D + F)(y + \delta y) = c + \delta c,$$

où

$$F = {}'P\delta A'Q, \quad y = Qx, \quad \delta y = Q\delta x, \quad c = {}'Pb, \quad \delta c = {}'P\delta b.$$

Notons que pour la norme spectrale ou euclidienne on a $\|F\| = \|\delta A\|$. Admettons que cette norme est petite ou qu'on a au moins $\|\delta A\| \cdot \|A^{-1}\| = \|F\| \cdot \|D^{-1}\| < 1$. Il vient

$$D(1 + D^{-1}F)(y + \delta y) = c + \delta c,$$

$$y + \delta y = (1 - D^{-1}F + \ldots)D^{-1}(c + \delta c).$$

Si l'on néglige les carrés des perturbations, on obtient en utilisant $y = D^{-1}c$ que

$$\delta y = -D^{-1}FD^{-1} + D^{-1}\delta c.$$

La matrice $D^{-1}FD^{-1}$ s'obtient de F par division de toutes ses lignes et colonnes par les nombres singuliers correspondants. Ceci étant, les éléments diagonaux se divisent par les carrés des nombres singuliers. Le dernier terme est obtenu par division des composantes de δc par les nombres singuliers correspondants.

Si la matrice A est quasi singulière et mal conditionnée, quelques-uns de ses nombres singuliers (au moins α_n) sont petits, tandis que les autres (α_1 en tout cas) ne peuvent être considérés comme petits. La dernière composante de δy est de la forme

$$\sum_{j=1}^{n-1} \frac{f_{nj}c_j}{\alpha_j\alpha_n} + \frac{f_{nn}c_n}{\alpha_n^2} + \frac{\delta c_n}{\alpha_n}.$$

Si l'on admet que f_{ij} sont comparables en grandeur avec α_n, le terme $f_{nn}c_n/\alpha_n^2$ dépassera en module tous les autres termes ainsi que les autres composantes qui ne contiennent que le terme $f_{in}c_n/\alpha_i\alpha_n$.

Il va de soi que si quelques autres nombres singuliers de la matrice A s'avèrent aussi petits, le résultat obtenu sera également vrai pour les composantes de l'erreur δy qui leur sont associées.

On a posé $\delta y = Q\delta x$. Les colonnes de la matrice Q constituent la seconde base singulière de la matrice A ou, ce qui revient au même, la première base singulière de la matrice A^{-1} (proposition 14, § 1, ch. XI). Aussi l'estimation assez grossière faite plus haut permet-elle de justifier la proposition suivante.

PROPOSITION 5. *Supposons que les nombres singuliers de la matrice A sont rangés en deux groupes dont l'un contient tous les nombres proches de zéro et l'autre, ceux qui en sont loin. Dans ce cas, l'erreur de la solution du système d'équations linéaires à matrice A est due aux composantes de A suivant les vecteurs de la première base singulière de la matrice A^{-1}, qui correspondent aux grands nombres singuliers de la matrice A^{-1}.*

L'interprétation géométrique de cette proposition peut être obtenue à l'aide de l'exemple étudié au § 1. Le vecteur de la première base singulière de la matrice A^{-1}, correspondant au plus grand nombre singulier est dirigée le long du plus grand axe de l'ellipse construite.

On peut de même fournir l'explication intuitive suivante de ce résultat. Si la matrice régulière A tend vers la matrice singulière $\hat{A}$, certains de ses nombres singuliers tendent vers zéro. Les vecteurs correspondants de la seconde base singulière de A tendent alors vers les vecteurs qui engendrent le sous-espace des solutions du système d'équations homogènes $\hat{A}\boldsymbol{\eta} = \mathbf{0}$.

Si $\boldsymbol{\xi}$ est solution du système $\hat{A}\boldsymbol{\xi} = \boldsymbol{b}$, toute matrice-colonne $\boldsymbol{\xi} + \boldsymbol{\eta}$, où $\hat{A}\boldsymbol{\eta} = \mathbf{0}$, est une solution de ce système, c'est-à-dire que la solution est définie à l'« erreur » près vérifiant le système homogène.

Terminons l'étude des matrices quasi singulières par une remarque particulière qui sera utile à l'exposé ultérieur. Pour une matrice symétrique, les modules des nombres caractéristiques sont des nombres singuliers, de sorte que si α_1 et α_n sont le plus grand et le plus petit nombre singulier de la matrice A, le plus grand et le plus petit nombre singulier de la matrice ${}^tAA$ seront α_1^2 et α_n^2. Cela signifie que pour le nombre conditionnel spectral on a

$$c({}^tAA) = [c(A)]^2. \qquad (13)$$

En outre, il peut arriver que pour un α_n assez élevé, α_n^2 acquiert un ordre comparable à celui des erreurs et, par suite, la matrice ${}^tAA$ sera quasi singulière, bien que A ne le soit pas.

4. Conditionnement du problème de recherche des vecteurs propres et des valeurs propres. Soit la transformation linéaire **A** définie dans une certaine base par la matrice A. On s'intéresse aux valeurs propres et aux vecteurs propres de cette transformation. Supposons qu'il nous soit donné, au lieu de la matrice A, la matrice perturbée $A + \delta A$. Il existe plusieurs résultats qui aident à estimer la perturbation des vecteurs propres et des valeurs propres, engendrée par la perturbation δA. On étudiera certains d'entre eux.

Admettons que la transformation **A** est définie par une matrice de structure simple. En vertu de cette hypothèse, il existe une matrice S telle que $D = S^{-1}AS$ est une matrice diagonale :

$$D = \text{diag}\,(\lambda_1, \ldots, \lambda_n).$$

Dans ce cas,

$$S^{-1}(A + \delta A)S = D + S^{-1}\delta AS.$$

Supposons que λ^* est une valeur propre perturbée, c'est-à-dire qu'il existe une matrice-colonne $\boldsymbol{\xi}^* \neq \mathbf{0}$ telle que $(A + \delta A)\boldsymbol{\xi}^* = \lambda^*\boldsymbol{\xi}^*$. Cela signifie que

$$(D + S^{-1}\delta AS)\boldsymbol{\eta}^* = \lambda^*\boldsymbol{\eta}^*,$$

où $\boldsymbol{\eta}^* = S^{-1}\boldsymbol{\xi}^*$, ou

$$(D - \lambda^*E)\boldsymbol{\eta}^* = -(S^{-1}\delta AS)\boldsymbol{\eta}^*.$$

On a

$$\frac{\|(D - \lambda^*E)\boldsymbol{\eta}^*\|}{\|\boldsymbol{\eta}^*\|} \geqslant \inf_{\|\boldsymbol{\eta}\| = 1} \|(D - \lambda^*E)\boldsymbol{\eta}\| = \min |\lambda_i - \lambda^*|$$

et

$$\frac{\|(S^{-1}\delta AS)\boldsymbol{\eta}^*\|}{\|\boldsymbol{\eta}^*\|} \leqslant \sup_{\|\boldsymbol{\eta}\| = 1} \|(S^{-1}\delta AS)\boldsymbol{\eta}\| = \|S^{-1}\delta AS\| \leqslant \|\delta A\|\, c(S),$$

où $c(S)$ est le nombre conditionnel spectral de la matrice S. Donc

$$\min_i |\lambda_i - \lambda^*| \leqslant \|\delta A\|\, c(S).$$

La matrice S est évidemment définie d'une façon moins restrictive. Considérons le nombre

$$\nu(A) = \inf c(S)$$

qui est la borne inférieure des nombres conditionnels de toutes les matrices diagonalisant A. Il est évident que cette borne inférieure existe et $\nu(A) \geqslant 1$. On peut maintenant écrire

$$\min_i |\lambda_i - \lambda^*| \leqslant \nu(A)\|\delta A\|. \tag{14}$$

Proposition 6. *La valeur propre perturbée λ^* appartient à l'un au moins des disques du plan complexe de centres aux points λ_i et de rayon $\nu(A)\delta$, où δ est la norme spectrale de la perturbation de la matrice.*

On peut montrer, comme pour les disques de localisation, que la réunion de m disques (14) contient exactement m valeurs propres perturbées.

Ceci étant, il faut tenir compte des multiplicités des valeurs propres perturbées et non perturbées.

Ainsi donc, le conditionnement du problème de recherche des valeurs propres peut, à certain égard, être décrit par le nombre $\nu(A)$. Les colonnes de la matrice S peuvent être considérées comme normées. Un grand nombre conditionnel signifiera dans ce cas que les colonnes de S sont proches des colonnes linéairement dépendantes. Donc, un $\nu(A)$ grand montre que A n'a pas de bonne base des vecteurs propres « suffisamment linéairement indépendants ».

Toutefois, on ne peut se limiter à un seul nombre conditionnel dans le problème de recherche des valeurs propres, car la situation est ici plus compliquée que dans le cas de résolution d'un système d'équations linéaires. Plus précisément, les valeurs propres différentes ne sont pas également sensibles aux perturbations de la matrice.

Soit $x = \| x_1, \ldots, x_n \|$ une base de vecteurs propres de la transformation **A**. Supposons que les vecteurs sont normés : $\| x_i \| = 1$ (on choisit la norme euclidienne en admettant que l'espace est euclidien). On sait de la proposition 10 du § 1, ch. XI, que la base biorthogonale à la base x est composée des vecteurs propres de la transformation adjointe $\mathbf{A}^*$ de **A**. Normons les vecteurs de cette base et notons les vecteurs normés $y_1, \ldots, y_n$.

Soient $\boldsymbol{\xi}_i$ et $\boldsymbol{\eta}_i$ les colonnes de coordonnées des vecteurs x_i et y_i, $i = 1, \ldots, n$, dans une base orthonormée $\boldsymbol{e}$. Dans ce cas, $A\boldsymbol{\xi}_i = \lambda_i \boldsymbol{\xi}_i$, ${}'A\boldsymbol{\eta}^i = \lambda_i \boldsymbol{\eta}_i$, ou ce qui revient au même ${}'\boldsymbol{\eta}_i A = \lambda_i {}'\boldsymbol{\eta}_i$ et, de plus,

$$ {}'\boldsymbol{\eta}_j \boldsymbol{\xi}_i = \begin{cases} 0, & i \neq j, \\ s_i \neq 0, & i = j. \end{cases} $$

Considérons le vecteur propre x^* et la valeur propre λ^* d'une transformation dont la matrice est perturbée. La colonne de coordonnées $\boldsymbol{\xi}^*$ du vecteur x^* vérifie l'égalité

$$ (A + \delta A)\boldsymbol{\xi}^* = \lambda^* \boldsymbol{\xi}^*. \tag{15} $$

Soit λ_i la valeur propre de **A** la plus proche de λ^* (ou l'une des plus proches si elles sont équidistantes de λ^*). Désignons $\lambda^* - \lambda_i$ par $\delta\lambda_i$. Pour une transformation de structure simple, l'espace se décompose en somme directe de sous-espaces propres. Cette décomposition définit la projection de x^* sur le sous-espace propre correspondant à λ_i. Désignons cette projection par x_i, et $x^* - x_i$ par δx_i. En cas de besoin, en multipliant x^* par un facteur numérique, on peut considérer que $\| x_i \| = 1$ et inclure x_i dans la base. Ecrivons (15) par l'intermédiaire des perturbations :

$$ (A + \delta A)(\boldsymbol{\xi}_i + \delta\boldsymbol{\xi}_i) = (\lambda_i + \delta\lambda_i)(\boldsymbol{\xi}_i + \delta\boldsymbol{\xi}_i), $$

où $\boldsymbol{\xi}_i$ et $\delta\boldsymbol{\xi}_i$ sont les colonnes de coordonnées de x_i et δx_i. Simplifions cette

relation en négligeant les produits de perturbations. Il vient

$$\delta A \boldsymbol{\xi}_i + A \delta \boldsymbol{\xi}_i = \lambda_i \delta \boldsymbol{\xi}_i + \delta\lambda_i \boldsymbol{\xi}_i . \tag{16}$$

En multipliant les deux membres de cette égalité par la ligne ${}'\boldsymbol{\eta}_i$, on obtient

$${}'\boldsymbol{\eta}_i \delta A \boldsymbol{\xi}_i + {}'\boldsymbol{\eta}_i A \delta \boldsymbol{\xi}_i = \lambda_i {}'\boldsymbol{\eta}_i \delta \boldsymbol{\xi}_i + \delta\lambda_i {}'\boldsymbol{\eta}_i \boldsymbol{\xi}_i ,$$

d'où

$$\delta\lambda_i = \frac{{}'\boldsymbol{\eta}_i \delta A \boldsymbol{\xi}_i}{s_i}$$

et

$$|\delta\lambda_i| \leqslant \frac{\|\delta A\|}{|s_i|} .$$

Ainsi donc, si $|s_i|$ n'est pas trop petit, $|\delta\lambda_i|$ est comparable à $\|\delta A\|$. Pour un $|s_i|$ petit, la sensibilité de la valeur propre λ_i aux perturbations de la matrice devient importante.

En multipliant les deux membres de l'égalité (16) par ${}'\boldsymbol{\eta}_j$ pour $j \neq i$, on obtient

$${}'\boldsymbol{\eta}_j \delta A \boldsymbol{\xi}_i + \lambda_j {}'\boldsymbol{\eta}_j \delta \boldsymbol{\xi}_i = \lambda_i {}'\boldsymbol{\eta}_j \delta \boldsymbol{\xi}_i .$$

d'où, si λ_j est différent de λ_i,

$${}'\boldsymbol{\eta}_j \delta \boldsymbol{\xi}_i = \frac{{}'\boldsymbol{\eta}_j \delta A \boldsymbol{\xi}_i}{\lambda_i - \lambda_j} .$$

Les produits du premier membre ne diffèrent des composantes correspondantes de la perturbation δx_i dans la base $\| x_1, \ldots, x_n \|$ que par les facteurs s_j. En effet, soit

$$\delta x_i = \alpha_1 x_1 + \ldots + \alpha_n x_n .$$

Dans ce cas,

$${}'\boldsymbol{\eta}_j \delta \boldsymbol{\xi}_i = \alpha_j {}'\boldsymbol{\eta}_j \boldsymbol{\xi}_j = \alpha_j s_j .$$

Notons J_i l'ensemble des numéros j pour lesquels $\lambda_j \neq \lambda_i$. Alors pour tous les $j \in J_i$ on a

$$\alpha_j = \frac{{}'\boldsymbol{\eta}_j \delta A \boldsymbol{\xi}_i}{(\lambda_i - \lambda_j) s_j} .$$

La perturbation δx_i a été définie de manière que ses composantes suivant les vecteurs x_j soient égales à zéro pour $j \notin J_i$. Par suite,

$$\delta \boldsymbol{\xi}_i = \sum_{j \in J_i} \frac{{}'\boldsymbol{\eta}_j \delta A \boldsymbol{\xi}_i}{(\lambda_i - \lambda_j) s_j} \boldsymbol{\xi}_j , \tag{17}$$

et sa norme ne dépasse pas

$$\|\delta A\| \sum_{j \in J_i} (|\lambda_i - \lambda_j| |s_j|)^{-1}.$$

Cela montre que le vecteur propre associé à λ_i est plus sensible aux perturbations de la matrice si λ_i est proche des autres valeurs propres.

L'expression (17) contient les facteurs s_j correspondant aux valeurs propres qui *diffèrent* de λ_i. Leur petitesse peut témoigner de la grande valeur de la norme de δx_i. Mais il peut s'avérer que le vecteur δx_i est petit en norme, tandis que sont grandes certaines de ses composantes dans la base $\|x_1, \ldots, x_n\|$ car les vecteurs de base sont proches des vecteurs linéairement dépendants.

Le coefficient s_j est égal au cosinus de l'angle formé par les vecteurs η_j et ξ_j. Pour les transformations symétriques, tous les s_j atteignent leur valeur maximale 1. La grandeur $|s_j^{-1}|$ est appelée ***coefficient de gauchissement*** de la transformation **A**.

Etudions le lien entre les coefficients de gauchissement et le nombre $\nu(A)$. Connaissant les vecteurs x_i, on est en mesure de construire une matrice S diagonalisant la matrice A. Considérons la matrice S dont les colonnes sont $\xi_i/\sqrt{|s_i|}$. Alors, les lignes de sa matrice inverse seront ${}^t\eta_j/\sqrt{|s_j|}$. Pour la matrice S on a

$$c_E(S) = \left(\sum_{i=1}^{n} |s_i^{-1}|\right)^{1/2} \left(\sum_{j=1}^{n} |s_j^{-1}|\right)^{1/2} = \sum_{i=1}^{n} |s_i^{-1}|.$$

Or

$$\nu(A) \leqslant c(S) \leqslant c_E(S).$$

Donc,

$$\nu(A) \leqslant \sum_{i=1}^{n} |s_i^{-1}|. \tag{18}$$

Si toutes les valeurs propres sont distinctes, l'orientation de chaque vecteur propre est définie au sens du vecteur près, de sorte que les coefficients de gauchissement sont déterminés de façon univoque. Mais si une transformation de structure simple possède des valeurs propres multiples, les coefficients de gauchissement dépendent de la base choisie.

Considérons, par exemple, une transformation symétrique dans un espace tridimensionnel, possédant les valeurs propres 1, 2, 2. Dans une base orthonormée de vecteurs propres, tous les coefficients de gauchissement possèdent leur valeur minimale 1. Mais si l'on fait tourner l'un des vecteurs correspondant à $\lambda = 2$ de l'angle $\pi/3$ dans son sous-espace bidimensionnel, il lui correspondra le coefficient de gauchissement égal à 2.

Pour éviter l'influence de cette non-univocité, dans le cas des racines multiples il faut étudier les bornes inférieures des coefficients de gauchissement suivant toutes les bases. Il va de soi que cela complique fortement leur utilisation.

Lors de la déduction de la formule (18), on partait d'une base de vecteurs propres arbitraire, si bien que l'inégalité (18) est également vérifiée pour les bornes inférieures des coefficients de gauchissement.

Le résultat suivant n'est vrai que pour les transformations qui n'admettent pas de valeurs propres multiples. A savoir, pour tout i,

$$|s_i^{-1}| \leqslant \nu(A). \tag{19}$$

En effet, étant donné une matrice S diagonalisant A, on peut choisir les bases $\|x_1, \ldots, x_n\|$ et $\|y_1, \ldots, y_n\|$ en posant

$$\xi_i = \frac{Se_i}{\|Se_i\|}, \quad {}^t\eta_i = \frac{{}^te_iS^{-1}}{\|{}^te_iS^{-1}\|},$$

où e_i est une colonne de la matrice unité. D'où

$$\|s_i^{-1}\| = \|Se_i\| \cdot \|{}^te_iS^{-1}\| \leqslant \|S\| \cdot \|S^{-1}\| = c(S).$$

Si les coefficients de gauchissement sont déterminés de façon univoque et la relation précédente est vérifiée pour toute matrice S, il s'ensuit l'assertion nécessaire.

§ 3. Méthodes directes de résolution des systèmes d'équations linéaires

On décrira dans ce paragraphe les principales méthodes directes de résolution des systèmes d'équations linéaires. Ces méthodes conduisent théoriquement à une solution exacte du système et par suite s'appellent aussi méthodes exactes à la différence des méthodes itératives fournissant en principe une solution approchée. Les méthodes itératives seront étudiées au § 5.

Soit donné un système d'équations linéaires de la forme

$$Ax = b, \tag{1}$$

où A est une matrice carrée régulière d'ordre n. Par matrice régulière on entend ici et plus loin une matrice qui n'est pas quasi singulière au sens indiqué au § 2. Aucune structure spéciale de la matrice A n'est supposée, vu que les méthodes directes sont généralement utilisées pour résoudre les systèmes dont les matrices peuvent être enregistrées toutes entières dans la mémoire vive de l'ordinateur. On parlera en premier lieu de la résolution des systèmes d'équations linéaires mais les mêmes transformations de

matrices peuvent être utilisées et sont utilisées pour le calcul des déterminants et des matrices inverses. On le signalera par des remarques appropriées.

Les méthodes de calcul sont appréciées suivant trois qualités principales : a) le nombre d'opérations arithmétiques effectuées, qui définit le temps de calcul ; b) l'exigence envers la capacité de la mémoire vive et c) la précision maximale atteinte par ces méthodes. Bien que la confrontation des méthodes soit une affaire délicate car son résultat dépend de circonstances variées et est souvent de nature arbitraire, on peut dégager deux groupes de méthodes qui sont les meilleurs. Les méthodes du premier groupe peuvent être réunies sous l'appellation commune de méthode de Gauss, celles du second groupe sont liées à la multiplication de la matrice du système par des matrices de transformations orthogonales. Ce sont ces deux groupes de méthodes qui feront l'objet de notre étude ultérieure.

1. Méthode de Gauss. D'une façon générale, la méthode de Gauss sert à transformer une matrice A en matrice unité par des opérations élémentaires sur les lignes de A. Si la matrice A est une matrice complète (contenant la matrice-colonne des seconds membres), la dernière colonne devient la solution du système.

Il existe plusieurs suites de transformations élémentaires pour passer de la matrice donnée à la matrice unité, ce qui engendre toute une série d'algorithmes réalisant la méthode de Gauss. L'un d'eux a été utilisé au § 4 du ch. V pour réduire la matrice à la forme simplifiée. Passons maintenant à l'algorithme appelé *schéma de division unique*.

Ce schéma comprend une série d'étapes successives où la matrice initiale $A = A^{(0)}$ se transforme en matrices $A^{(1)}, A^{(2)}, \ldots$ Commençons par le plus simple des cas quand tous les mineurs principaux de la matrice A, c'est-à-dire les mineurs de la forme

$$\det\left\|\begin{matrix} a_{11} & \cdots & a_{1k} \\ \cdots & \cdots & \cdots \\ a_{k1} & \cdots & a_{kk} \end{matrix}\right\|, \quad k = 1, \ldots, n,$$

sont différents de zéro, ce qui permet d'omettre les opérations élémentaires de permutation des lignes et des colonnes.

Au cours de la première étape du schéma, la matrice A se transforme en matrice $A^{(1)}$ dont la première colonne est celle de la matrice unité d'ordre n. Par hypothèse, le mineur du premier ordre, égal à l'élément a_{11}, est différent de zéro. En divisant la première ligne de A par a_{11} (division unique), on transforme la matrice A en A', où $a'_{11} = 1$ et $a'_{k1} = a_{k1}$ pour tous les $k \geqslant 2$. Ensuite, pour $k = 2, \ldots, n$, on retranche la première ligne multipliée par a_{k1} de la k-ième ligne dans la matrice A'. On obtient ainsi la

matrice de la forme

$$A^{(1)} = \left\| \begin{matrix} 1 & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \dots & a_{2n}^{(1)} \\ \dots & \dots & \dots & \dots \\ 0 & a_{n2}^{(1)} & \dots & a_{nn}^{(1)} \end{matrix} \right\| .$$

Notons que dans les transformations décrites on ajoute une ligne à des lignes situées au-dessous d'elle et, par suite, les mineurs principaux de la matrice ne peuvent s'annuler. Ainsi donc, les mineurs principaux de la matrice $A^{(1)}$ sont différents de zéro. Il en découle que l'élément $a_{22}^{(1)}$ de cette matrice est non nul car le mineur principal d'ordre 2 vaut $a_{22}^{(1)}$.

La deuxième étape du schéma de division unique consiste en application des transformations de la première étape à une sous-matrice de la matrice $A^{(1)}$, qui est formée des lignes et des colonnes de numéros 2, ..., n. On obtient ainsi la matrice

$$A^{(2)} = \left\| \begin{matrix} 1 & a_{12}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & 1 & \dots & a_{2n}^{(2)} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn}^{(2)} \end{matrix} \right\|$$

dont les deux premières colonnes ont les unités sur la diagonale principale et les zéros au-dessous de la diagonale.

Dans les transformations effectuées, une ligne s'ajoutait de nouveau à celles d'au-dessous et, par suite, les mineurs principaux de la matrice $A^{(2)}$ ne sont pas nuls. En particulier, est différent de zéro l'élément $a_{33}^{(2)}$ qui est égal au mineur principal d'ordre 3 de la matrice $A^{(2)}$.

A la troisième étape, on transforme la sous-matrice de la matrice $A^{(2)}$, formée des lignes et des colonnes de numéros 3, ..., n. Et ainsi de suite. A la k-ième étape, les transformations de la première étape sont appliquées à la sous-matrice de la matrice $A^{(k-1)}$, formée des $n - k + 1$ dernières lignes et colonnes. Dans la matrice $A^{(k)}$ obtenue, l'élément $a_{k+1,\,k+1}^{(k)}$ est différent de zéro, sinon s'annulerait par des transformations élémentaires des lignes le mineur principal d'ordre $k + 1$ de la matrice A.

Après la $(n - 1)$-ième étape on obtient la matrice $A^{(n-1)}$, dont tous les éléments sur la diagonale principale, sauf $a_{nn}^{(n-1)}$, sont égaux à 1, et les éléments situés au-dessous de la diagonale principales sont nuls.

Les matrices dont les éléments situés au-dessous de la diagonale sont nuls sont appelées *matrices triangulaires supérieures* (voir p. 307). En divisant la dernière ligne de la matrice $A^{(n-1)}$ par $a_{nn}^{(n-1)}$, on obtient la matrice triangulaire supérieure $A^{(n)} = U$ avec des unités sur la diagonale principale.

Le processus, décrit plus haut, de transformation de la matrice A par des opérations élémentaires en matrice triangulaire supérieure U porte le nom de ***méthode d'élimination directe*** ou de ***méthode de Gauss*** au sens restreint du terme.

Si on effectue toutes les opérations élémentaires sur les lignes de la matrice complète, on obtient un système d'quations linéaires à matrice triangulaire

$$Ux = b',$$

qui est équivalent au système (1) (voir proposition 2, § 4, ch. V). Il peut être résolu facilement. En effet, la dernière équation du système est de la forme $x_n = b_n'$. De plus, quel que soit k, si $x_n, \ldots, x_k$ sont définis, la $(k - 1)$-ième équation entraîne

$$x_{k-1} = b_{k-1}' - \sum_{j=k}^{n} u_{k-1,j} x_j. \tag{2}$$

En appliquant cette formule successivement pour $k = n, \ldots, 2$, on obtient toutes les composantes de la solution sans recourir à la division.

Le procédé décrit de résolution du système à matrice triangulaire est appelé ***méthode de substitution inverse.***

Au lieu d'effectuer une substitution inverse, on peut transformer U en matrice unité à l'aide des opérations élémentaires sur les lignes, analogues à celles qui transforment A en U. A cet effet retranchons la n-ième ligne multipliée par des facteurs convenables de toutes les lignes situées au-dessus d'elle, de sorte que tous les éléments de la n-ième colonne situés au-dessus de la diagonale s'annulent. Ensuite, en nous servant de la $(n - 1)$-ième ligne, annulons les éléments de la $(n - 1)$-ième colonne situés au-dessus de la diagonale, etc. Ce processus de transformation est appelé ***marche*** (ou ***opération***) ***inverse*** de la méthode de Gauss.

Si toutes les opérations sont effectuées sur les lignes d'une matrice complète, le système sera remplacé, à la suite d'opérations directe et inverse de la méthode de Gauss, par un système équivalent à matrice unité, c'est-à-dire sera résolu.

Notons que l'opération inverse de la méthode de Gauss exige un plus grand nombre d'opérations arithmétiques que la résolution du système à matrice triangulaire par les formules (2).

On reviendra plus loin dans le point 3 au schéma de division unique en décrivant comment se libérer de l'hypothèse restrictive que tous les mineurs principaux de la matrice sont différents de zéro. Maintenant, on étudiera sommairement la méthode dite d'***élimination optimale*** qui servira d'exemple du degré de modification des caractéristiques de l'algorithme avec la variation de l'ordre des opérations effectuées.

Le processus d'élimination optimale débute comme le schéma de division unique par la division de tous les éléments de la première ligne par a_{11}. (On admet toujours que tous les mineurs principaux de la matrice sont différents de zéro.) La nouvelle première ligne est multipliée par a_{21} et retranchée de la deuxième. Toutefois, on ne procède pas à la transformation de la troisième ligne, mais à l'aide de la deuxième on annule le premier élément dans la deuxième colonne et c'est seulement alors qu'on passe à la troisième ligne.

Dans la troisième ligne, à l'aide des deux premières, on annule les deux premiers éléments et, à l'aide de la troisième ligne, on annule les deux premiers éléments de la troisième colonne. Sans entrer dans les détails, posons qu'après k étapes les k premières lignes sont transformées de manière que les parties des k premières colonnes qui y sont contenues se confondent avec les colonnes de la matrice unité d'ordre k. A la $(k + 1)$-ième étape on ajoute à la $(k + 1)$-ième ligne les lignes de numéros $1, \ldots, k$ multipliées par des facteurs convenables pour annuler ses k premiers éléments. Ensuite, on divise la $(k + 1)$-ième ligne par son $(k + 1)$-ième élément et, après la multiplication par les éléments adéquats, on la retranche des lignes situées au-dessus d'elle, de manière à annuler les k premiers éléments de la $(k + 1)$-ième colonne.

Ainsi, dans la méthode d'élimination optimale on procède tour à tour aux opérations directe et inverse de la méthode de Gauss. Cela permet de ne pas introduire dans la mémoire vive de l'ordinateur la $(k + 1)$-ième ligne avant que ne soient transformées les lignes précédentes.

Après la transformation de k lignes, le nombre d'éléments à mémoriser diminue dans ces lignes de k^2 et devient égal à $nk - k^2$. La valeur maximale de cette expression est atteinte, si n est pair, pour $k = n/2$ et vaut $n^2/4$. Pour un n impair, le résultat a une valeur proche. Donc, la résolution du système à l'aide de la méthode d'élimination optimale exige une capacité de la mémoire vive à peu près quatre fois moindre qu'avec l'utilisation du schéma de division unique.

2. ***LU*-décomposition.** On sait qu'une transformation élémentaire quelconque des lignes de la matrice A équivaut à la multiplication de A à gauche par une matrice régulière, et qu'une suite de telles transformations, à la multiplication par la matrice S égale au produit de matrices correspondantes. Pour obtenir la matrice S, il suffit d'effectuer successivement toutes les opérations élémentaires sur la matrice unité. (En effet, on obtient ainsi la matrice SE, c'est-à-dire S.) En transformant la matrice A en matrice triangulaire supérieure U, on effectue une suite déterminée d'opérations élémentaires. On voit aussitôt que cette suite d'opérations transforme E en une matrice triangulaire inférieure S. En effet, outre la multiplication des lignes par des nombres, on ne se sert que de l'addition d'une ligne à des

lignes situées au-dessous d'elle. Dans ce cas, tous les éléments de la matrice unité transformée, se trouvant au-dessus de la diagonale restent nuls. Ainsi donc, on a la proposition suivante.

PROPOSITION 1. *Pour toute matrice A dont les mineurs principaux sont non nuls, il existe une matrice triangulaire inférieure régulière S telle que SA est la matrice triangulaire supérieure U avec des unités sur la diagonale principale.*

Démontrons ensuite la

PROPOSITION 2. *La matrice L inverse de la matrice triangulaire inférieure S de la proposition* 1 *est elle-même une matrice triangulaire inférieure de la forme*

$$L = \left\| \begin{matrix} a_{11} & 0 & \dots & 0 & \dots & 0 \\ a_{21} & a_{22}^{(1)} & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & a_{kk}^{(k-1)} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2}^{(1)} & \dots & a_{nk}^{(k-1)} & \dots & a_{nn}^{(n-1)} \end{matrix} \right\|,$$

où $a_{mk}^{(k-1)}$ sont des éléments de la matrice $A^{(k-1)}$ obtenue suivant le schéma de division unique à la $(k-1)$-ième étape.

Pour le démontrer, considérons les transformations élémentaires du schéma de division unique comme résultat de multiplication par la matrice. Notons $S^{(k)}$ une matrice telle que

$$A^{(k)} = S^{(k)}A^{(k-1)}.$$

A la k-ième étape, on divise la k-ième ligne de la matrice $A^{(k-1)}$ par $a_{kk}^{(k-1)}$ et on soustrait la ligne obtenue multipliée par $a_{mk}^{(k-1)}$ de toutes les lignes de numéros $m = k + 1, \dots, n$. La matrice $S^{(k)}$ est obtenue à partir de la matrice unité à l'aide des mêmes transformations élémentaires et, par suite, n'en diffère que par les éléments de la k-ième colonne, qui se trouvent sur ou sous la diagonale :

$$S^{(k)} = \left\| \begin{matrix} 1 & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & s_k^{(k)} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & s_n^{(k)} & \dots & 1 \end{matrix} \right\|.$$

Ici $s_k^{(k)} = (a_{kk}^{(k-1)})^{-1}$ et $s_m^{(k)} = -a_{mk}^{(k-1)}/a_{kk}^{(k-1)}$ pour $m > k$.

Il nous faut démontrer que

$$S^{(n)} \dots S^{(2)}S^{(1)}L = E.$$

Considérons le produit $S^{(1)}L$. La première colonne de L coïncide avec la première colonne de A, et la multiplication à gauche par $S^{(1)}$ transforme la première colonne de A en première colonne e_1 de la matrice unité, de sorte que la première colonne de la matrice $S^{(1)}L$ se confond avec e_1. Les autres colonnes de la matrice $S^{(1)}L$ sont les mêmes que dans la matrice L. En effet, la multiplication par $S^{(1)}$ est équivalente à l'addition de la première ligne multipliée par des facteurs appropriés à des lignes situées au-dessous d'elle ; quant aux éléments de la première ligne de L, ils sont nuls à partir du second.

Ensuite, la multiplication à gauche par $S^{(2)}$ est équivalente à l'addition de la deuxième ligne multipliée par des facteurs convenables à des lignes se trouvant au-dessous d'elle. Dans la matrice $S^{(1)}L$, tous les éléments de la deuxième ligne, sauf $a_{22}^{(1)}$, sont nuls. Donc $S^{(2)}S^{(1)}L$ ne diffère de $S^{(1)}L$ que par la deuxième colonne. Les éléments de la deuxième colonne de la matrice $S^{(1)}L$ situés sur ou sous la diagonale coïncident avec les éléments homologues de la matrice $A^{(1)}$. Par suite, la multiplication par $S^{(2)}$ transforme la deuxième colonne de la matrice $S^{(1)}L$ en deuxième colonne de la matrice unité.

En continuant de raisonner ainsi pour toutes les autres matrices $S^{(i)}$, on aboutit à l'assertion nécessaire.

Il est très important que les éléments de la k-ième colonne de la matrice L sont des éléments de la matrice $A^{(k-1)}$. Si la matrice A ne doit pas être conservée ou si elle est stockée dans la mémoire externe, on peut, après chaque transformation élémentaire, enregistrer les éléments de la matrice obtenue à l'endroit des éléments correspondants de la matrice qui a subi une transformation. Dans ce cas, les éléments de la matrice L peuvent être enregistrés à la place des zéros (au-dessous de la diagonale) ou des unités (sur la diagonale) dans la matrice transformée. Après toutes les transformations, on obtient au-dessus de la diagonale les éléments de la matrice U qui ne sont pas *a priori* des zéros ou des unités, les autres places étant réservées aux éléments de la matrice L.

DÉFINITION. La décomposition de la matrice A en produit LU d'une matrice triangulaire inférieure régulière L et d'une matrice triangulaire supérieure U avec des unités sur la diagonale principale est appelée *LU-décomposition* *) de la matrice A.

Il découle des propositions 1 et 2 le théorème suivant.

THÉORÈME 1. *La matrice A admet la LU-décomposition si tous ses mineurs principaux (le déterminant y compris) sont différents de zéro.*

---

*) D'après les premières lettres des mots anglais « upper » et « lower » signifiant « supérieur » et « inférieur ».

On obtient la LU-décomposition en indiquant un algorithme qui permet de la calculer. Ceci étant, on n'a pas prêté attention à deux circonstances. On le fera à l'aide de la proposition suivante qui contient d'ailleurs une autre démonstration de l'existence de cette décomposition.

PROPOSITION 3. *Pour que la LU-décomposition de la matrice A existe, il est nécessaire* (*et suffisant*) *que les mineurs principaux de A soient non nuls. Si la LU-décomposition existe, elle est unique.*

DÉMONSTRATION. Soient $S = L^{-1}$ et $U = SA$. Notons U_k, S_k et A_k les sous-matrices des matrices U, S et A, qui sont situées à l'intersection des lignes et des colonnes de numéros 1, ..., k. Les éléments de U_k sont les produits des lignes de S par les colonnes de A, les numéros des lignes et des colonnes ne dépassant pas k. Or S est une matrice triangulaire inférieure. Donc, ces produits sont les mêmes que ceux des lignes de S_k par les colonnes de A_k. D'où on a $U_k = S_k A_k$.

Vu que det $U_k \neq 0$, on a aussi det $A_k \neq 0$, et par suite, les mineurs principaux de A doivent nécessairement être différents de zéro pour que la LU-décomposition existe.

On sait que la dernière ligne du produit est une combinaison linéaire des lignes du second facteur, dont les coefficients sont égaux aux éléments de la dernière ligne du premier facteur. La dernière ligne de U_k est la dernière ligne de la matrice unité et par suite, est parfaitement définie. Les lignes de A_k sont linéairement indépendantes, de sorte que chaque ligne à k éléments se décompose suivant ses dernières de façon unique. Donc, la k-ième ligne de S_k est parfaitement définie.

Or on obtient la k-ième ligne de S à partir de la k-ième ligne de S_k en lui joignant $n - k$ zéros. Vu que k est arbitraire, il s'ensuit que la matrice S est parfaitement définie. Ensuite, L se définit de façon unique en tant que S^{-1}, et U en tant que SA.

REMARQUE. Soulignons que n'est unique que la décomposition dont la deuxième matrice triangulaire possède des unités sur la diagonale principale. Il existe en général plusieurs décompositions en matrices triangulaires, en particulier, celle dont la première matrice possède des unités sur la diagonale principale.

La matrice L peut être représentée comme produit de la matrice L_1 avec des unités sur la diagonale principale et de la matrice diagonale D. On a alors la

PROPOSITION 4. *La matrice A dont les mineurs principaux ne sont pas nuls peut être décomposée d'une façon unique en produit $L_1 DU$, où D est une matrice diagonale et L_1 et U sont des matrices triangulaires inférieure et supérieure avec des unités sur la diagonale principale.*

L'unicité de la décomposition mentionnée découle de l'unicité de la

LU-décomposition, vu que la représentation $L = L_1 D$ est évidemment aussi unique.

Considérons la LDU-décomposition, obtenue dans la proposition 4, pour le cas particulier d'une matrice symétrique A. Alors $A = {}^tA = {}^tU\,D\,{}^tL$, où ${}^tU$ est la matrice triangulaire inférieure et ${}^tL$ la matrice triangulaire supérieure, toutes deux avec des unités sur la diagonale principale. En vertu de l'unicité de la décomposition, on a $L = {}^tU$ et

$$A = {}^tUDU. \tag{3}$$

En interprétant la matrice A comme une matrice associée à la forme quadratique, on peut considérer que l'égalité (3) exprime le passage à une base dans laquelle la forme quadratique se définit par une matrice diagonale D.

En particulier, si tous les mineurs principaux de la matrice A sont strictement positifs, la forme quadratique est définie positive et tous les éléments diagonaux de la matrice D sont strictement positifs. Si $D = \operatorname{diag}(d_1, \ldots$ $\ldots, d_n)$, où tous les $d_i > 0$, on peut introduire la matrice $D^{1/2} = \operatorname{diag}(\sqrt{d_1}, \ldots$ $\ldots, \sqrt{d_n})$ et la matrice $V = D^{1/2}U$. Il vient alors

$$A = {}^tVV. \tag{4}$$

Il existe un algorithme efficace permettant d'obtenir directement la décomposition (4) pour une matrice A définie positive. On le décrira au point 5.

Notons que la décomposition (4) n'est pas un fait inattendu. On peut considérer toute matrice définie positive comme la matrice de Gram d'une base $\boldsymbol{e}$ pour un produit scalaire convenablement défini. La formule (4) peut alors être interprétée comme la relation entre les matrices de Gram de deux bases, à savoir : la base $\boldsymbol{e}$ et la base orthonormée par rapport au produit scalaire considéré, la matrice de passage étant une matrice triangulaire supérieure. La matrice inverse de cette dernière, c'est-à-dire la matrice de passage de $\boldsymbol{e}$ à une base orthonormée, est aussi une matrice triangulaire. La construction de la matrice de passage d'une base quelconque à une base orthonormée est réalisée par orthogonalisation de Gram-Schmidt. Il est aisé de vérifier que ce processus aboutit à une base orthonormée dont le k-ième vecteur est la combinaison linéaire des vecteurs $e_1, \ldots, e_k$ de la base $\boldsymbol{e}$. Ainsi, le processus d'orthogonalisation de Gram-Schmidt conduit justement à la matrice triangulaire supérieure qui est inverse de la matrice V dans la décomposition (4).

Revenons au sujet principal de ce point pour souligner l'importance du résultat obtenu. La LU-décomposition joue un grand rôle dans les méthodes numériques de résolution des systèmes d'équations linéaires, de calculs des déterminants et d'inversion des matrices. En effet, une grande partie

des difficultés liées à la résolution des problèmes mentionnés peut être rapportée à la recherche de la LU-décomposition. Si la matrice du système d'équations linéaires se présente sous forme de produit LU, la résolution du système se réduit à la résolution successive de deux systèmes à matrices triangulaires. A savoir, $Ax = b$ équivaut à $Ly = b$ et $Ux = y$.

En notant ensuite que $\det A = \det L \det U$ et $\det U = 1$, on peut calculer $\det A$ comme le produit des éléments diagonaux de la matrice L.

De même, $A = LU$ est équivalent à $A^{-1} = U^{-1}L^{-1}$ et, vu que les matrices triangulaires peuvent facilement être inversées, le calcul de A^{-1} ne présente pas de difficulté.

Le mérite non des moindres de la LU-décomposition est qu'elle occupe dans la mémoire vive de l'ordinateur autant de place que la matrice initiale.

3. Choix d'un élément principal. Considérons maintenant une matrice régulière A sans imposer aucune restriction à ses mineurs principaux. Il se peut que la condition $a_{11} \neq 0$ ne soit plus remplie. Dans ce cas, pour transformer la première colonne de la matrice A en colonne de la matrice unité il ne suffit plus d'un nombre fini d'opérations élémentaires utilisées dans le schéma de division unique, décrit plus haut. Toutefois, si l'on ajoute à ces opérations les permutations des lignes (ou des colonnes), on peut ne plus exiger dans le théorème 1 que les mineurs principaux soient non nuls.

En effet, on peut énoncer la proposition suivante.

PROPOSITION 5. *Toute matrice régulière A d'ordre n peut être transformée par permutation des seules lignes (ou des seules colonnes) en une matrice dont les mineurs principaux sont différents de zéro.*

Les $n - 1$ premières colonnes de la matrice A contiennent nécessairement un mineur différent de zéro, car autrement les colonnes seraient linéairement dépendantes et $\det A$ serait nul. Permutons les lignes de manière que ce mineur devienne le mineur principal d'ordre $n - 1$. Soit A' la matrice obtenue par la permutation réalisée. Il va de soi que dans A' on peut permuter les lignes de numéros $1, \ldots, n - 1$ sans annuler son mineur principal d'ordre $n - 1$. Profitons-en pour placer sur la diagonale principale le mineur non nul d'ordre $n - 2$ situé quelque part sur les $n - 1$ premières lignes et les $n - 2$ premières colonnes. Un tel mineur doit obligatoirement exister, car autrement le mineur principal d'ordre $n - 1$ serait nul dans la matrice A'.

En continuant d'agir de la sorte avec les mineurs d'ordres décroissants, on démontrera l'assertion pour les permutations des lignes. Pour les permutations des colonnes la démonstration est analogue.

La permutation des lignes dans une matrice est équivalente à la multiplication à gauche de cette matrice par la matrice P obtenue de la matrice unité E par la même permutation des lignes. D'une façon analogue, la per-

mutation des colonnes est équivalente à la multiplication à droite par la matrice Q obtenue de E par la même permutation des colonnes. Les matrices P et Q sont appelées *matrices de permutation.*

La matrice inverse de la matrice de permutation est une matrice associée à la permutation inverse qui retourne les lignes à leurs anciennes places. Vu que la matrice de permutation est orthogonale, on a pour elle $P^{-1} = {}^tP$.

En interprétant les permutations mentionnées dans la proposition 5 comme multiplication par une matrice, on aboutit, en vertu des résultats obtenus auparavant sur la LU-décomposition, au théorème suivant.

THÉORÈME 2. *Pour toute matrice régulière A on a les décompositions suivantes :*

$$A = PLU, \qquad A = L'U'Q,$$
$$A = PMDV, \qquad A = M'D'VQ,$$

où P et Q sont des matrices de permutation, U, U', V et V' des matrices triangulaires supérieures avec des unités sur la diagonale principale, L, L', M et M' des matrices triangulaires inférieures telles que M et M' possèdent des unités sur la diagonale, et D, D' sont des matrices diagonales.

Pour appliquer ce résultat, il faut indiquer un algorithme permettant de construire les matrices de permutation ou, ce qui revient au même, indiquer la permutation des lignes ou des colonnes rendant les mineurs principaux différents de zéro. A la place de cet algorithme, on recourt en général au schéma de division unique sous forme modifiée. Ce schéma modifié aboutit à une suite de matrices de transformation où alternent des matrices triangulaires et des matrices de permutation. Les modifications sont les suivantes.

S'il arrive qu'au début de la première étape l'élément a_{11} est nul, on permute les lignes de manière qu'à la place de a_{11} vienne un élément non nul. Cet élément existe obligatoirement dans la première colonne car det $A \neq 0$. On dira que c'est un élément *principal* (ou clé) de la première étape.

De même, si à la deuxième étape on a $a_{22}^{(1)} = 0$, la deuxième colonne de la matrice $A^{(1)}$ contient un élément non nul situé au-dessous de la diagonale (car autrement les deux premières colonnes seraient linéairement dépendantes). En permutant les lignes, on peut placer cet élément à la place de $a_{22}^{(1)}$. Il s'appellera élément principal de la deuxième étape. De façon analogue on détermine un élément principal de chaque étape.

On a décrit le choix d'un élément principal suivant la colonne. Si on remplace les permutations des lignes par celles des colonnes, l'élément principal de la k-ième étape sera sélectionné parmi les éléments de la k-ième ligne de la matrice $A^{(k-1)}$. Le choix d'un élément principal suivant la matrice entière signifie qu'il est possible de mettre à la place de $a_{kk}^{(k-1)}$ tout

élément $a_{ij}^{(k-1)}$ pour $i > k$, $j > k$. Notons que le choix d'un élément principal suivant la ligne conduit à la deuxième décomposition dans le théorème 2.

Théoriquement, à titre d'élément principal on peut prendre tout élément non nul. Toutefois en pratique la situation n'est pas si simple. Comme on l'a vu au § 1, même le résultat de vérification de l'égalité à zéro d'un élément peut dépendre de la représentation des nombres dans la machine à calculer. Le choix de l'élément principal influe de façon sensible sur les erreurs d'arrondi. Aussi, même si l'on a établi que par exemple a_{11} n'est pas nul dans la matrice initiale, cela ne signifie-t-il nullement qu'il est rationnel de le choisir pour élément principal. Considérons à titre d'exemple le système

$$\begin{aligned} 10^{-4}x + y &= 1, \\ x + y &= 2 \end{aligned}$$

et admettons que les résultats des opérations arithmétiques sont arrondis à trois chiffres significatifs dans le système à virgule flottante. La solution du système calculée sans arrondi est :

$$\begin{aligned} x &= (1 - 10^{-4})^{-1} = 1 + 10^{-4} + 10^{-8} + \ldots, \\ y &= 2 - x = 1 - 10^{-4} - 10^{-8} - \ldots \end{aligned}$$

En adoptant 10^{-4} pour élément principal, on doit transformer la matrice complète du système de la façon suivante :

$$\left\| \begin{array}{cc|c} 10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array} \right\| \to \left\| \begin{array}{cc|c} 1 & 10^4 & 10^4 \\ 1 & 1 & 2 \end{array} \right\| \to \left\| \begin{array}{cc|c} 1 & 10^4 & 10^4 \\ 0 & -10^4 & -10^4 \end{array} \right\| \to \left\| \begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 1 \end{array} \right\|.$$

Notons que le nombre 2, après qu'on y ait ajouté 10^4 et arrondi le résultat, a disparu, de sorte que si dans le système initial on avait 3 et non pas 2, le résultat serait le même. La solution trouvée est $x = 0$, $y = 1$, ce qui est loin de la solution exacte.

Mais si après avoir permuter les lignes on choisit 1 pour élément principal, la transformation de la matrice complète sera :

$$\left\| \begin{array}{cc|c} 10^{-4} & 1 & 1 \\ 1 & 1 & 2 \end{array} \right\| \to \left\| \begin{array}{cc|c} 1 & 1 & 2 \\ 10^{-4} & 1 & 1 \end{array} \right\| \to \left\| \begin{array}{cc|c} 1 & 1 & 2 \\ 0 & 1 & 1 \end{array} \right\| \to \left\| \begin{array}{cc|c} 1 & 0 & 1 \\ 0 & 1 & 1 \end{array} \right\|,$$

ce qui nous donne la solution $x = 1$, $y = 1$ qui coïncide avec la solution exacte arrondie, c'est-à-dire une solution aussi précise qu'il est en général possible.

Dans la première variante, on a vu que la perte de précision était due à la nécessité d'additionner les nombres dont les ordres diffèrent de plus que de la longueur de la mantisse. Ceci étant, le plus petit nombre disparaît. Même pour une différence d'ordres inférieure, une partie de chiffres signi-

ficatifs du plus petit terme se perd. Aussi, pour diminuer les erreurs d'arrondi, faut-il s'arranger à additionner les nombres dont les ordres sont voisins.

Généralement, on recommande de choisir pour élément principal un élément dont le module est maximal dans la partie de la matrice à transformer ou, ce qui est moins bon, bien que plus simple, un élément de module maximal de la ligne ou colonne considérée. Au point qui suivra on reviendra à cette question. Maintenant, remarquons seulement qu'en multipliant, dans l'exemple étudié plus haut, la première équation par 10^5, on obtient le système

$$\begin{aligned} 10x + 10^5 y &= 10^5, \\ x + y &= 2. \end{aligned}$$

L'élément maximal de la première colonne est 10. En le choisissant pour élément principal, on obtient encore une solution peu satisfaisante.

Dans certains cas, la recherche d'une solution plus précise ne détermine pas à elle seule le choix de l'élément principal. Si on applique la méthode de Gauss à une matrice creuse (voir p. 382), chaque opération élémentaire fait croître en général le nombre d'éléments non nuls dans la matrice. Il peut donc arriver qu'on ne puisse enregistrer l'une des matrices transformées dans la partie accessible de la mémoire. Dans ce cas, on n'obtiendra pas de solution exacte car on n'obtiendra aucune solution. Toutefois, le nombre d'éléments non nuls de la matrice peut croître de façon différente suivant le choix de l'élément principal. Aussi l'approche suivante est-elle possible.

On peut estimer *a priori* la variation du nombre d'éléments non nuls de la matrice pour les suites différentes d'éléments principaux choisis et s'arrêter sur la suite qui donne la valeur minimale (ou admissible) de ce nombre. Cette méthode est décrite dans le livre de Twearson [37] déjà mentionné. Il va de soi que le procédé proposé est assez laborieux mais il faut prendre en compte qu'il est souvent nécessaire de résoudre plusieurs systèmes d'équations linéaires ayant une même matrice des coefficients (et ne différant que par les seconds membres) ou des matrices qui présentent la même disposition des éléments non nuls. Dans ce cas, la peine qu'on se donne pour choisir une suite d'éléments principaux peut s'avérer justifiée.

4. Mise à l'échelle. La sélection de l'élément principal est intimement liée à la multiplication des équations (ou ce qui revient au même, des lignes de la matrice du système) par des facteurs numériques. Cette transformation équivaut à la multiplication à gauche par une matrice diagonale et est appelée *mise à l'échelle des lignes*. La multiplication des colonnes par des nombres (*mise à l'échelle des colonnes*) équivaut à la multiplication à droite de la matrice du système par une matrice diagonale et peut être interprétée comme un passage à d'autres unités de mesure pour les inconnues. D'où son appellation.

Pour facteurs de mise à l'échelle il est commode de choisir les puissances de la base du système de numération (de dix par exemple, si le système est décimal). Dans l'arithmétique à virgule flottante, ces facteurs ne modifient que les ordres des éléments en conservant les mantisses, et par suite, la mise à l'échelle n'introduit pas d'erreurs d'arrondi supplémentaires.

Si le choix de l'élément principal est déterminé par une règle liée à la grandeur des éléments de la matrice, il dépend donc de la mise à l'échelle. Par exemple, si pour élément principal de la matrice on choisit le plus grand élément en module, une mise à l'échelle convenable peut rendre principal tout élément non nul donné *a priori*. En effet, par la mise à l'echelle des lignes on peut rendre chaque élément non nul de la i_0-ième ligne plus grand en module que tout élément d'une autre ligne. Ensuite, par la mise à l'échelle des colonnes, on peut rendre maximal en module l'élément situé à l'intersection de la i_0-ième ligne et de la j_0-ième colonne, cet élément choisi restant supérieur aux autres éléments de la j_0-ième colonne.

Les mêmes considérations peuvent être aussi appliquées à la sélection de l'élément principal suivant une ligne ou une colonne.

Mais si le choix de l'élément principal se détermine par une règle ne dépendant pas de la grandeur des éléments de la matrice, par exemple, si la position des éléments principaux est définie *a priori*, la mise à l'échelle ne change pas la solution. Plus précisément, on aboutit au résultat suivant.

Supposons pour fixer les idées qu'on utilise l'arithmétique décimale à virgule flottante, et considérons deux systèmes d'équations linéaires dont les matrices A et A' diffèrent, par la mise à l'échelle de leurs lignes et colonnes, de puissances entières de dix, c'est-à-dire que leurs éléments sont liés par les relations

$$a'_{ij} = 10^{p_i+q_j}a_{ij}, \quad i, j = 1, \ldots, n,$$

tandis que les seconds membres des deux systèmes satisfont aux conditions

$$b'_i = 10^{p_i}b_i, \quad i = 1, \ldots, n.$$

On peut alors énoncer la proposition qui suit (voir Forsythe, Moler [10]).

Proposition 6. *Supposons que les deux systèmes se résolvent suivant le schéma de division unique et que dans les deux systèmes on choisit à chaque étape pour éléments principaux les éléments qui occupent les mêmes positions dans les matrices. Dans ce cas, les solutions calculées des deux systèmes sont liées par l'égalité*

$$x'_j = 10^{-q_j}x_j, \quad j = 1, \ldots, n, \tag{5}$$

et les mantisses de x'_j et de x_j coïncident si la résolution d'un des systèmes n'entraîne pas le zéro de machine ou un dépassement de capacité.

(On dit qu'on obtient la solution du second système à partir de celle du premier par la « démise à l'échelle ».)

DÉMONSTRATION. Notons d'abord que les mantisses des éléments correspondants de deux systèmes envisagés se confondent et que l'arrondissement d'un nombre marque sa mantisse et non son ordre qui peut entraîner l'apparition d'un dépassement de capacité ou du zéro de machine.

Vérifions que l'application d'une même opération élémentaire du schéma de division unique aux deux matrices fournit les mêmes systèmes ne différant que par la mise à l'échelle. Pour simplifier l'écriture, posons qu'il s'agit de la première étape dans l'opération directe du schéma de division unique.

Admettons que pour élément principal est choisi un élément situé à l'intersection de la k-ième ligne et de la l-ième colonne. Dans la matrice A, de la i-ième ligne on retranche le produit de la k-ième ligne par a_{il}/a_{kl}. Dans ce cas, le j-ième élément de la ligne obtenue est égal à

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{il}}{a_{kl}} a_{kj}.$$

Dans la matrice A', une transformation analogue donne

$$a_{ij}'^{(1)} = a_{ij} 10^{p_i+q_j} - \frac{a_{il} 10^{p_i+q_l}}{a_{kl} 10^{p_k+q_l}} a_{kj} 10^{p_k+q_j} = 10^{p_i+q_j} a_{ij}^{(1)}.$$

Pour les éléments de la k-ième ligne dans la matrice A' transformée on a

$$a_{kj}'^{(1)} = \frac{a'_{kj}}{a'_{kl}} = \frac{a_{kj} 10^{p_k+q_j}}{a_{kl} 10^{p_k+q_l}} = 10^{-q_l+q_j} a_{kj}^{(1)}.$$

En particulier, $a_{kl}'^{(1)} = a_{kl}^{(1)} = 1$ pour $j = l$. Dans la matrice-colonne b' des termes constants on retranche du i-ième élément le k-ième multiplié par a'_{il}/a'_{kl} :

$$b_i'^{(1)} = b_i 10^{p_i} - b_k 10^{p_k} \frac{a_{il} 10^{p_i+q_l}}{a_{kl} 10^{p_k+q_l}} = b_i^{(1)} 10^{p_i}.$$

De plus, le k-ième élément de la matrice-colonne des termes constants doit être divisé par a'_{kl} :

$$b_k'^{(1)} = \frac{b_k 10^{p_k}}{a_{kl} 10^{p_k+q_l}} = b_k^{(1)} 10^{-q_l}.$$

Ainsi donc, les deux systèmes transformés sont liés de la même façon que les systèmes initiaux, et les facteurs de mise à l'échelle n'ont pas varié, à l'exception du facteur de la k-ième ligne qu'on a remplacé par 10^{-q_l} du fait de la division par a'_{kl}. Les opérations directe et inverse étant réalisées,

on aboutit aux systèmes à matrices unités (à condition de faire toutes les permutations nécessaires). Dans les opérations directe et inverse, chaque ligne a rempli son rôle de ligne clé une seule fois, et lors de l'opération inverse on n'a pas effectué la division. C'est pourquoi le facteur de mise à l'échelle de la ligne contenant l'unité à la l-ième place vaut 10^{-q_l}. On aboutit ainsi à l'expression (5). La proposition est démontrée.

Les raisonnements faits montrent que si on convient de choisir pour élément principal l'élément maximal de la colonne, on ne fixe pas pour le système donné une suite d'éléments principaux mais on établit une relation entre les mises à l'échelle et les suites d'éléments principaux. Avec cette règle de sélection d'éléments principaux, on ne doit pas ignorer la mise à l'échelle : sans introduire les facteurs de mise à l'échelle on conserve celle-ci telle qu'elle était, ce qui n'est peut-être pas sa meilleure variante.

A ce qu'il paraît, la meilleure mise à l'échelle doit être celle pour laquelle le nombre conditionnel de la matrice est minimal. Or les algorithmes efficaces au choix de telle mise à l'échelle ne sont pas connus. On se sert ainsi de l'approche suivante.

Soit une norme $\| * \|$ sur l'espace des matrices-colonnes à n éléments. Une matrice d'ordre n est dite *équipondérante en colonnes* (resp. *lignes*) par rapport à la norme choisie si chacune de ses colonnes (resp. lignes) vérifie la condition

$$0{,}1 \leqslant \|a_i\| \leqslant 1$$

(pour le système de numération décimale). La matrice est dite *équipondérante* si elle est équipondérante en lignes et colonnes.

Avant de résoudre un système d'équations linéaires, on peut mettre à l'échelle sa matrice de façon qu'elle devienne équipondérante. Le défaut de ce procédé est qu'il peut engendrer plusieurs matrices équipondérantes conditionnées très différemment.

5. Calculs avec double précision et schéma compact. Toute une série d'ordinateurs présente des possibilités techniques, traduites dans les langages de programmation qui y sont adoptés (Fortran IV, PL/1), d'effectuer tous les calculs avec un nombre double de chiffres, ce qui augmente évidemment, de façon sensible, la précision du résultat. On ne s'arrêtera pas sur la discussion de ces possibilités car on n'a nulle part fixé le nombre t de chiffres significatifs et son doublement n'apportera rien de nouveau à nos raisonnements. Toutefois, l'analyse des erreurs d'arrondi montre que tous les calculs effectués suivant un algorithme quelconque n'ont pas la même influence sur la précision du résultat. La précision de tout le processus de calcul peut être souvent fortement améliorée du fait de l'accroissement de la précision d'une seule partie de ces calculs.

Pour une large classe de calculateurs, il existe une possibilité d'obtenir,

avec une grande précision et d'une façon relativement simple, sans presque augmenter le temps de calcul, les expressions de la forme $\Sigma \alpha_k \beta_k$, c'est-à-dire les produits de lignes par des colonnes. C'est ce qu'on appelle *calcul du produit scalaire en régime d'accumulation*. Voyons plus en détail de quoi il s'agit.

Supposons qu'on doit multiplier deux nombres dans le système à virgule flottante. Il faut pour cela additionner leurs ordres et multiplier les mantisses. Si les mantisses sont des nombres décimaux à t chiffres, leur produit exact aura $2t$ chiffres. Le produit des mantisses est supérieur à 0,01, mais il peut être inférieur à 0,1. Dans ce dernier cas, on doit le multiplier par 10 et oter 1 de l'ordre. Après quoi, le produit des mantisses est arrondi jusqu'à t chiffres. Le nombre obtenu sera la mantisse du produit, tandis que les chiffres rejettés définissent l'erreur d'arrondi.

Toutefois, si l'on a besoin de la somme de produits, on peut se passer de l'arrondi et, au fur et à mesure du calcul de produits, les additionner à la somme des produits déjà obtenus en les interprétant comme des nombres à $2t$ chiffres significatifs dans le système à virgule flottante. Dans ce cas, on n'arrondit jusqu'à t chiffres significatifs qu'après avoir additionné le dernier produit. La somme de produits ainsi calculée diffère par rapport à sa valeur exacte d'une grandeur qui dépasse très peu cette erreur d'arrondi, à condition qu'il ne se produise dans l'addition une grande perte de chiffres significatifs due à la présence des termes dont les ordres diffèrent sensiblement. La démonstration de ce fait peut être trouvée dans le livre de Forsythe et Moler [10].

Le calcul de la somme de produits figure pratiquement dans tous les algorithmes de résolution des problèmes d'algèbre linéaire, si bien que la possibilité de calculer de telles sommes avec de plus faibles erreurs d'arrondi est très importante. Fixons l'attention sur le fait que la formule (2) par exemple permet de résoudre le système d'équations linéaires à matrice triangulaire par le calcul de la somme de produits. Ainsi donc, en utilisant le régime d'accumulation, on obtient une solution du système qui contient des erreurs d'arrondi comparables à l'erreur d'arrondi produite par une seule multiplication.

La possibilité de calculer la somme de produits en régime d'accumulation a engendré la création d'une série d'algorithmes permettant de mieux utiliser cette possibilité. Etudions un algorithme de recherche de la LU-décomposition, appelé *schéma compact* de la méthode de Gauss.

L'égalité matricielle $A = LU$ est équivalente à n^2 égalités numériques

$$a_{ij} = \sum_{k=1}^{n} l_{ik} u_{kj}.$$

Vu que L et U sont des matrices triangulaires, on a $l_{ik} = 0$ pour $k > i$, et $u_{kj} = 0$ pour $k > j$. Donc,

$$a_{ij} = \sum_{k=1}^{r} l_{ik} u_{kj}, \tag{6}$$

où $r = \min(i, j)$. Ces n^2 égalités peuvent être groupées suivant les valeurs de r. Ainsi, pour $r = 1$ on a ou $j > i = 1$, ou $i \geqslant j = 1$, de sorte que le groupe défini par $r = 1$ est constitué des équations

$$\begin{aligned} a_{1j} &= l_{11} u_{1j} \quad & j > 1, \\ a_{i1} &= l_{i1} u_{11} \quad & i \geqslant j = 1. \end{aligned}$$

Vu que U possède des unités sur la diagonale principale, $u_{11} = 1$ et l'on obtient

$$l_{i1} = a_{i1} \text{ pour tous les } i \geqslant 1.$$

En particulier, $l_{11} = a_{11}$ et, par suite, $l_{11} \neq 0$ si les mineurs principaux de la matrice A sont différents de zéro. On peut maintenant trouver

$$u_{1j} = \frac{a_{1j}}{l_{11}} \text{ pour tous les } j > 1.$$

Admettons maintenant que dans les groupes d'équations correspondant aux valeurs de $r = 1, \ldots, s - 1$, on a trouvé l_{ir} et u_{rj} pour tous i et j et tous les $r \leqslant s - 1$. Ecrivons le groupe d'équations correspondant à $r = s$ sous la forme

$$a_{ij} = l_{is} u_{sj} + \sum_{k=1}^{s-1} l_{ik} u_{kj}.$$

Il s'ensuit pour $i \geqslant j = s$, en vertu de $u_{ss} = 1$, que

$$l_{is} = a_{is} - \sum_{k=1}^{s-1} l_{ik} u_{ks}.$$

La somme dans le second membre de l'égalité se compose des éléments déjà connus, et donc on a trouvé l_{is} pour $i \geqslant s$. Pour $i < s$, posons $l_{is} = 0$.

Si $j > i = s$, on a

$$l_{ss} u_{sj} = a_{sj} - \sum_{k=1}^{s-1} l_{sk} u_{kj}.$$

Démontrons que $l_{ss} \neq 0$. En effet, s'il n'en est pas ainsi, ou bien n'existe aucun nombre u_{sj} vérifiant la dernière égalité, ou bien il en existe une infinité (suivant que le premier membre de l'égalité est nul ou non). Dans les

deux cas, on aboutit à la contradiction avec la proposition 3. Donc,

$$u_{sj} = \frac{a_{sj} - \sum_{k=1}^{s-1} l_{sk} u_{kj}}{l_{ss}} \quad \text{pour tous les } j > s.$$

Pour $j < s$ posons $u_{sj} = 0$. Ainsi donc, en utilisant l'équation (6) pour $r = 1, \dots, s$, on a obtenu tous les l_{is} et u_{sj}, ce qui montre que tous les éléments des matrices L et U peuvent être calculés successivement avec utilisation du régime d'accumulation. On peut montrer que parmi toutes les méthodes connues le schéma compact décrit ici de la méthode de Gauss présente la meilleure majoration pour la norme de la perturbation des données initiales, quand cette perturbation est équivalente à l'influence des erreurs d'arrondi. Le nombre d'opérations arithmétiques réalisées d'après ce schéma, ainsi que la capacité nécessaire de la mémoire sont à peu près les mêmes que ceux du schéma de division unique, c'est-à-dire sont proches du minimum possible.

Remarque. On a montré plus haut que les éléments diagonaux par lesquels on doit effectuer les divisions dans le schéma compact sont différents de zéro. Or cela ne garantit qu'une possibilité théorique de la réalisation des calculs. Pour que ces calculs deviennent pratiquement réalisables et stables par rapport aux perturbations des données initiales et aux erreurs d'arrondi, il faut que les nombres l_{ss} ne soient pas petits ou, ce qui revient au même, que les nombres u_{sj} ne soient pas grands. Pour ces derniers on a $u_{sj} = a_{sj}^{(s)}$, où $a_{sj}^{(s)}$ est un élément de la matrice $A^{(s)}$ construite dans le schéma de division unique. La même condition, absence de grands éléments dans la matrice $A^{(s)}$, apparaît aussi dans l'étude de la stabilité du schéma de division unique.

On a étudié ici l'utilisation du schéma compact à la recherche de la LU-décomposition, c'est-à-dire admis que les mineurs diagonaux de la matrice A sont différents de zéro. Il faut noter que le schéma compact possède le défaut fondamental de la méthode de Gauss : la stabilité du processus de calcul est fonction du choix des éléments principaux ou, dans le cas considéré, de l'ordre dans lequel se disposent les lignes et les colonnes de la matrice avant le calcul.

On a montré plus haut que toute matrice symétrique définie positive admet une décomposition de la forme $A = V'V$, où V est la matrice triangulaire inférieure. La modification du schéma compact de recherche de cette décomposition s'appelle *méthode de la racine carrée*. La décomposition cherchée, exprimée par les éléments des matrices, est de la forme

$$a_{ij} = \sum_{k=1}^{r} v_{ik} v_{jk},$$

où $r = \min(i, j)$. De même que les équations (6), ces équations peuvent être résolues successivement. En effet, pour $i = j = 1$ et $i > j = 1$ on a respectivement

$$a_{11} = v_{11}^2 \quad \text{et} \quad a_{i1} = v_{i1}v_{11},$$

d'où $v_{11} = \sqrt{a_{11}}$, $v_{i1} = a_{i1}/\sqrt{a_{11}}$.

Ensuite, si sont connus v_{ik} pour $k = 1, \ldots, s - 1$ et pour tous les i, l'égalité (pour $i = j = s$)

$$a_{ss} = v_{ss}^2 + \sum_{k=1}^{s-1} v_{sk}^2$$

permet de trouver v_{ss}, et l'égalité (pour $i > j = s$)

$$a_{is} = v_{is}v_{ss} + \sum_{k=1}^{s-1} v_{ik}v_{sk}$$

donne v_{is}. Toutes les divisions et les extractions de racine peuvent en principe être réalisées, vu que la décomposition cherchée, comme on l'a démontré, existe et est unique.

Comme le schéma compact, la méthode de la racine carrée admet le calcul du produit scalaire en régime d'accumulation. Ceci étant, on peut atteindre une très haute précision si la matrice est bien conditionnée. Dans le cas contraire où une matrice est mal conditionnée, les erreurs d'arrondi peuvent même violer la propriété de la matrice d'être définie positive et rendre ainsi impossible la réalisation de l'algorithme.

6. Décomposition en produit de matrices orthogonale et triangulaire. La décomposition de la matrice carrée A en produit de facteurs

$$A = QR, \tag{7}$$

où R est une matrice triangulaire supérieure et Q une matrice orthogonale joue un grand rôle dans les méthodes numériques de l'algèbre linéaire. On l'appelle *QR-décomposition* de A.

Si la matrice du système d'équations linéaires (1) admet la QR-décomposition, le système (1) se réduit au système

$$Rx = {}^tQb$$

à matrice triangulaire, après quoi il peut être résolu facilement par le procédé décrit à la p. 400.

L'application de la QR-décomposition à la résolution des systèmes d'équations linéaires est justifiée par le fait que le nombre conditionnel de la matrice R est égal à celui de la matrice A (comp. p. 386).

L'une des méthodes de recherche de la QR-décomposition porte le nom

de *méthode des symétries*. On appelle *symétrie* dans l'espace euclidien $\mathscr{E}_n$ la transformation symétrique possédant la valeur propre 1 de multiplicité $n - 1$ et la valeur propre -1 de multiplicité 1.

Si à la valeur propre -1 est associé le vecteur propre s de longueur 1, on dira que la symétrie est engendrée par le vecteur s.

La matrice de la symétrie par rapport à la base orthonormée $\boldsymbol{e}_0$ composée de ses vecteurs propres est de la forme

$$P_0 = \begin{Vmatrix} -1 & & & 0 \\ & 1 & & \\ & & \ddots & \\ 0 & & & 1 \end{Vmatrix} = E - 2E_{1,1}, \text{ avec } E_{1,1} = \begin{Vmatrix} 1 & & & 0 \\ & 0 & & \\ & & \ddots & \\ 0 & & & 0 \end{Vmatrix}.$$

La matrice P_0 est othogonale et, par suite, la symétrie est une transformation orthogonale, ce qui explique justement l'application des symétries à la résolution des systèmes d'équations linéaires.

Rapportée à une base orthonormée $\boldsymbol{e}$ déduite de $\boldsymbol{e}_0$ par une matrice de passage orthogonale S, la matrice de la symétrie prend la forme

$$P = S^{-1}P_0S = E - 2S^{-1}E_{1,1}S.$$

Désignons par ε_{kl} et σ_{ij} les éléments des matrices respectives $E_{1,1}$ et S. L'élément de la matrice $S^{-1}E_{1,1}S$ situé à l'intersection de la i-ième ligne et de la j-ième colonne est alors de la forme

$$\sum_{k,l} \sigma_{ki}\varepsilon_{kl}\sigma_{lj}$$

(il a été tenu compte de ce que $S^{-1} = {}^tS$). Tous les ε_{kl} sont égaux à 0, à l'exception de ε_{11} égal à 1, de sorte que la somme écrite plus haut vaut $\sigma_{1i}\sigma_{1j}$ et

$$S^{-1}E_{1,1}S = \boldsymbol{\sigma}{}^t\boldsymbol{\sigma},$$

où $\boldsymbol{\sigma}$ est la première colonne de la matrice ${}^tS$. Or ${}^tS$ est la matrice de passage de $\boldsymbol{e}$ à $\boldsymbol{e}_0$. La matrice-colonne $\boldsymbol{\sigma}$ est la colonne de coordonnées dans la base $\boldsymbol{e}$ du premier vecteur de la base $\boldsymbol{e}_0$, c'est-à-dire du vecteur s définissant la symétrie. On voit que dans une base orthonormée la matrice de la symétrie est de la forme

$$P = E - 2\boldsymbol{\sigma}{}^t\boldsymbol{\sigma}, \tag{8}$$

où $\boldsymbol{\sigma}$ est la colonne de coordonnées du vecteur définissant la symétrie.

PROPOSITION 7. *Quels que soient le vecteur non nul* x *et le vecteur* f *de longueur* 1, *il existe une symétrie que transforme* x *en* αf.

Notons dès le début qu'on a $|\alpha| = |x|$, car la symétrie est une transformation orthogonale. Pour fixer les idées, posons que $\alpha = |x|$. On démontrera que la symétrie nécessaire est définie par le vecteur unité s colinéaire à $x - \alpha f$.

A cet effet, introduisons une base orthonormée et notons $\boldsymbol{\xi}$ et $\boldsymbol{\varphi}$ les colonnes de coordonnées des vecteurs x et f. Déterminons la longueur λ du vecteur $x - \alpha f$:

$$|x - \alpha f|^2 = {}'(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi})(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi}) = {}'\boldsymbol{\xi}\boldsymbol{\xi} - 2\alpha\,{}'\boldsymbol{\xi}\boldsymbol{\varphi} + \alpha^2({}'\boldsymbol{\varphi})\boldsymbol{\varphi}.$$

Vu que $\alpha^2 = |x|^2 = {}'\boldsymbol{\xi}\boldsymbol{\xi}$ et ${}'\boldsymbol{\varphi}\boldsymbol{\varphi} = |f|^2 = 1$, il vient

$$\lambda^2 = |x - \alpha f|^2 = (2\alpha^2 - 2\alpha\,{}'\boldsymbol{\xi}\boldsymbol{\varphi}).$$

Considérons le vecteur s de colonne de coordonnées $\lambda^{-1}(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi})$ et la symétrie qu'il définit, de matrice

$$P = E - 2\lambda^{-2}(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi})\,{}'(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi}).$$

L'image du vecteur x par cette symétrie est

$$P\boldsymbol{\xi} = \boldsymbol{\xi} - 2\lambda^{-2}(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi})\,{}'(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi})\boldsymbol{\xi}.$$

Or

$${}'(\boldsymbol{\xi} - \alpha\boldsymbol{\varphi})\boldsymbol{\xi} = {}'\boldsymbol{\xi}\boldsymbol{\xi} - \alpha\,{}'\boldsymbol{\varphi}\boldsymbol{\xi} = \frac{1}{2}\lambda^2.$$

Donc, $P\boldsymbol{\xi} = \boldsymbol{\xi} - \boldsymbol{\xi} + \alpha\boldsymbol{\varphi}$ et l'on voit que la symétrie obtenue possède la propriété exigée. La proposition est démontrée.

Proposition 8. *La matrice carrée A peut être décomposée en produit QR, où Q est une matrice orthogonale et R une matrice triangulaire supérieure.*

Pour le démontrer, posons que l'espace des matrices-colonnes à n éléments est équivalent à l'espace euclidien rapporté à la base orthonormée donnée. En se servant de la proposition 7, on transformera successivement les colonnes de la matrice A.

Si la première colonne de A est nulle, passons à la colonne suivante. Dans le cas contraire, on cherche la symétrie qui transforme la première colonne de A en une colonne qui est proportionnelle à la première colonne de la matrice unité. Soit $P^{(1)}$ la matrice de cette symétrie. Alors

$$P^{(1)}A = A^{(1)},$$

et la matrice $A^{(1)}$ prend la forme

$$\left\|\begin{matrix} \alpha_1 & a_{12}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \ldots & a_{2n}^{(1)} \\ .. & \ldots\ldots & \ldots & \ldots\ldots \\ 0 & a_{n2}^{(1)} & \ldots & a_{nn}^{(1)} \end{matrix}\right\|$$

où l'élément α_1 est égal à la norme euclidienne de la première colonne dans la matrice A.

Supposons maintenant qu'après $k - 1$ étapes on a trouvé les matrices des symétries $P^{(1)}, \ldots, P^{(k-1)}$ telles que la matrice $A^{(k-1)} = P^{(k-1)} \ldots P^{(1)}A$ d'éléments $a_{ij}^{(k-1)}$ possède des zéros dans les $k - 1$ premières colonnes, au-dessous de la diagonale principale.

A la k-ième étape on construit la symétrie qui transforme la colonne

$$'\left\| \underbrace{0, \ldots, 0}_{k-1}, a_{kk}^{(k-1)}, \ldots, a_{nk}^{(k-1)} \right\| \tag{9}$$

en k-ième colonne de la matrice unité. Ces deux colonnes ont des zéros aux $k - 1$ premières places. Il en est donc de même de la colonne de coordonnées $\boldsymbol{\sigma}$ du vecteur qui engendre la symétrie. Il s'ensuit directement selon (8) que les $k - 1$ premières lignes et colonnes de la matrice associée à la symétrie ne diffèrent pas des lignes et colonnes homologues de la matrice unité, c'est-à-dire qu'elle est de la structure suivante

$$P^{(k)} = \left\| \begin{matrix} E_{k-1} & O \\ O & \bar{P} \end{matrix} \right\|. \tag{10}$$

Si les $n - k$ derniers éléments de la colonne (9) sont nuls, la symétrie ne peut être construite par le procédé décrit plus haut, car le vecteur $\boldsymbol{\sigma}$ devient nul. Mais dans ce cas, la colonne de la matrice A n'a pas besoin d'être transformée et la symétrie peut être omise.

Il est évident qu'en multipliant la matrice (10) par une colonne quelconque $\boldsymbol{\xi}$, on obtient la colonne $P^{(k)}\boldsymbol{\xi}$ dont les $k - 1$ premier éléments sont ceux de $\boldsymbol{\xi}$, tandis que les $n - k + 1$ derniers éléments sont indépendants des $k - 1$ premiers éléments de la colonne $\boldsymbol{\xi}$ et s'obtiennent par multiplication du bloc $\bar{P}$ par le tronçon inférieur de la colonne $\boldsymbol{\xi}$ qui contient $n - k + 1$ éléments.

On voit que les $k - 1$ premières lignes et colonnes de la matrice

$$A^{(k)} = P^{(k)}A^{(k-1)}$$

sont les mêmes que celles de la matrice $A^{(k-1)}$ et que de plus la k-ième colonne est de la forme

$$'\left\| a_{1k}^{(k-1)}, \ldots, a_{k-1,k}^{(k-1)}, \alpha_k, \underbrace{0, \ldots, 0}_{n-k} \right\|.$$

Ainsi, les éléments de la matrice $A^{(k)}$, situés au-dessous de la diagonale principale dans les k premières colonnes, sont nuls.

Notons que le produit des symétries est une transformation orthogonale qui n'est plus en général symétrique.

Après la $(n - 1)$-ième étape, on aboutit à la matrice triangulaire supé-

rieure $A^{(n-1)}$ qu'on notera R. Si $P^{(n-1)} \ldots P^{(1)} = P$ et $Q = {}^tP$, l'égalité $R = P^{(n-1)} \ldots P^{(1)} A$ est équivalente à la décomposition en facteurs (7).

PROPOSITION 9. *Si la matrice A est régulière, sa QR-décomposition dans laquelle les éléments diagonaux de R sont strictement positifs est unique.*

DÉMONSTRATION. Soit $A = Q_1 R_1 = Q_2 R_2$. Etant donné que la matrice A est régulière, il en est de même de la matrice R_1, et l'on peut écrire

$$ {}^tQ_2 Q_1 = R_2 R_1^{-1}. $$

Notons U la matrice $R_2 R_1^{-1}$. Etant le produit de deux matrices triangulaires supérieures, elle est aussi une matrice triangulaire supérieure. U^{-1} est également une matrice triangulaire supérieure. Mais en vertu de l'égalité précédente, U est une matrice orthogonale, donc $U^{-1} = {}^tU$ est une matrice triangulaire inférieure. Cela signifie que U est une matrice diagonale. Or la matrice orthogonale et diagonale ne peut avoir sur la diagonale que $+1$ ou -1. Etant donné que les éléments diagonaux des matrices R_1 et R_2 sont strictement positifs, il ressort de $R_2 = UR_1$ que $U = E$. La conclusion nécessaire est maintenant évidente.

La décomposition en matrices orthogonale et triangulaire, obtenue à l'aide de la méthode des symétries, peut être utilisée dans le calcul du déterminant et de la matrice inverse. En effet, $\det A = \det Q \det R$. Ici $\det Q = (-1)^s$, où s est le nombre de symétries effectuées, car le déterminant de la matrice d'une symétrie vaut -1. Le déterminant de la matrice R se calcule directement par multiplication de ses éléments diagonaux.

Pour trouver la matrice inverse, remarquons que $A = QR$ entraîne

$$ A^{-1} = R^{-1}({}^tQ) = R^{-1} P^{(n-1)} \ldots P^{(1)}, $$

et tout se réduit à l'inversion de la matrice triangulaire supérieure R. On a montré dans la proposition 2 comment on peut inverser une matrice triangulaire inférieure. La matrice triangulaire supérieure est inversée de façon analogue.

Estimons la capacité de mémoire nécessaire à l'enregistrement de la QR-décomposition obtenue par la méthode des symétries. Le produit de la matrice d'une symétrie par une colonne quelconque $\boldsymbol{\xi}$ peut être obtenu d'après la formule (8) :

$$ P\boldsymbol{\xi} = \boldsymbol{\xi} - 2\boldsymbol{\sigma}\,{}^t\boldsymbol{\sigma}\boldsymbol{\xi}. $$

Cela montre qu'il n'est pas nécessaire de mémoriser toutes les matrices des symétries. Il suffit de conserver les coordonnées des vecteurs engendrant ces symétries, notamment celles des coordonnées qui ne sont pas *a priori* nulles. Ces coordonnées sont pour les $n-1$ vecteurs au nombre de $n + (n-1) + \ldots + 2 = n(n+1)/2 - 1$. Pour mémoriser les éléments de la matrice R qui ne se trouvent pas au-dessous de la diagonale

principale, il faut réserver $n(n+1)/2$ cellules. Ainsi, pour mémoriser les deux matrices Q et R, il ne faut que $n-1$ cellules de plus que pour mémoriser la matrice A.

7. Méthode des rotations. Le second procédé largement utilisé à la recherche de la QR-décomposition est la *méthode des rotations*. Pour la décrire, considérons dans l'espace euclidien $\mathscr{E}_n$ une base orthonormée $\| e_1, \ldots, e_n \|$. On appelle *rotation* dans le plan des vecteurs e_i et e_j $(i<j)$ de cette base la transformation dont la matrice

$$P = \begin{Vmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & \cos\varphi & & & -\sin\varphi & & \\ & & & \ddots & & & & \\ & & & & 1 & & & \\ & & & & & \ddots & & \\ & & \sin\varphi & & & \cos\varphi & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{Vmatrix} \tag{11}$$

ne diffère de la matrice unité que par une sous-matrice d'ordre 2 située à l'intersection des lignes et colonnes de numéros i et j. Dans cette transformation, le plan des vecteurs e_i et e_j subit une rotation d'angle φ, tandis que chaque vecteur du supplémentaire orthogonal de ce plan demeure inchangé. Le produit d'une matrice de la forme (11) par une matrice-colonne $\boldsymbol{\xi}_0$ est une matrice-colonne dont tous les éléments sont ceux de $\boldsymbol{\xi}_0$ à l'exception des i-ième et j-ième qui se transforment suivant les formules

$$\begin{aligned} \xi_0'^{i} &= \xi_0^i \cos\varphi - \xi_0^j \sin\varphi, \\ \xi_0'^{j} &= \xi_0^i \sin\varphi + \xi_0^j \cos\varphi. \end{aligned} \tag{12}$$

Il est évident que quels que soient ξ_0^i et ξ_0^j, il existe une rotation pour laquelle $\xi_0'^{i} = \sqrt{(\xi^i)^2 + (\xi^j)^2}$ et $\xi_0'^{j} = 0$. Les coefficients dans les formules (12) doivent être, dans le cas général, égaux à

$$\cos\varphi = \xi_0^i/\rho, \quad \sin\varphi = -\xi_0^j/\rho,$$

où $\rho = \sqrt{(\xi_0^i)^2 + (\xi_0^j)^2}$. Si $\xi_0^j = 0$, on a $\cos\varphi = 1$, $\sin\varphi = 0$, c'est-à-dire que la matrice P est une matrice unité.

Pour une matrice carrée régulière A on est en mesure de construire une suite de matrices de rotation $P^{(1)}, \ldots, P^{(N)}$ de la forme (11) telle que la matrice $R = P^{(N)} \ldots P^{(1)}A$ soit une matrice triangulaire supérieure (avec éléments non nuls sur la diagonale principale). On voit aisément que cette construction est équivalente à la recherche de la décomposition (7).

La suite des matrices $P^{(1)}, \ldots, P^{(N)}$ contient pour chaque colonne de A

autant de matrices qu'il y a, dans cette colonne, d'éléments non nuls sous la diagonale principale. On posera $N = n(n - 1)/2$ car, si parmi les éléments situés sous la diagonale de la matrice A il existe des éléments nuls, les matrices de rotation correspondantes peuvent être prises pour des matrices unités.

Considérons la première colonne de la matrice A et choisissons une rotation de matrice $P^{(1)}$ dans le plan des vecteurs e_1 et e_2 telle que l'élément $a_{12}^{(1)}$ de la matrice $A^{(1)} = P^{(1)}A$ devienne nul. Ensuite, on effectue la rotation de matrice $P^{(2)}$ dans le plan des vecteurs e_1 et e_3, qui annule l'élément $a_{13}^{(2)}$ de la matrice $A^{(2)} = P^{(2)}A^{(1)}$, etc. La rotation dans le plan des vecteurs e_1 et e_j ne modifie pas les éléments de la colonne dont les numéros diffèrent de 1 et j et, par suite, les rotations ultérieures ne modifieront pas les éléments qui ont été annulés par les rotations précédentes. Après $n - 1$ rotations, on obtient la matrice $A^{(n-1)} = P^{(n-1)} \dots P^{(1)}A$ dont tous les éléments de la première colonne, à l'exception du premier, sont nuls.

Les rotations annulant les éléments de la deuxième colonne qui sont situés sous la diagonale sont choisies de la même manière dans les plans des vecteurs e_2, e_j pour tous les $j = 3, \dots, n$. Notons leurs matrices $P^{(n)}, \dots$ $\dots, P^{(2n-3)}$. En posant $i = 2$ et $j > 2$ dans les formules (12), on constate que dans ces rotations les éléments nuls de la première colonne demeurent inchangés.

De la même façon, les rotations dans les plans des vecteurs e_m, e_j pour tous les $j = m + 1, \dots, n$ permettent d'annuler les éléments situés au-dessous de la diagonale dans toute colonne de numéro $m \leqslant n - 1$. Ces rotations laissent invariants les éléments situés sous la diagonale dans les colonnes précédentes, et ils resteront nuls.

8. Utilisation du processus d'orthogonalisation. La QR-décomposition est intimement liée au processus d'orthogonalisation de Gram-Schmidt. Si les colonnes de la matrice A sont linéairement indépendantes, elles constituent une base dans l'espace arithmétique $\mathscr{R}^n$. Proposons-nous de les orthogonaliser. On sait que la première colonne $\boldsymbol{a}_1$ devient normée, c'est-à-dire est remplacée par $\boldsymbol{q}_1 = \rho_{11}\boldsymbol{a}_1$. La deuxième colonne $\boldsymbol{a}_2$ est remplacée par la combinaison linéaire $\boldsymbol{q}_2 = \rho_{12}\boldsymbol{a}_1 + \rho_{22}\boldsymbol{a}_2$ et, en général, la j-ième colonne doit être remplacée par

$$\boldsymbol{q}_j = \rho_{1j}\boldsymbol{a}_1 + \dots + \rho_{jj}\boldsymbol{a}_j. \tag{13}$$

Les coefficients ρ_{ij} sont choisis de manière que les colonnes $\boldsymbol{q}_j$ constituent un système orthonormé et, par suite, que la matrice $Q = \| \boldsymbol{q}_1, \dots, \boldsymbol{q}_n \|$ soit orthogonale. Considérons les matrices-colonnes

$$\boldsymbol{u}_j = {}^t\| \rho_{1j}, \dots, \rho_{jj}, 0, \dots, 0 \|$$

et la matrice U constituée avec ces dernières. U est une matrice triangulaire

supérieure et les égalités (13) signifient que $AU = Q$. Il est évident que cette égalité est équivalente à la QR-décomposition avec $R = U^{-1}$.

Ainsi, la méthode d'orthogonalisation de Gram-Schmidt peut être utilisée pour obtenir la QR-décomposition. Malheureusement, la stabilité numérique de cette méthode n'est pas grande. Les erreurs d'arrondi s'accumulent de façon qu'en fin de compte la matrice Q obtenue peut s'avérer absolument non orthogonale. Il existe des procédés pour améliorer ce processus, mais on les étudiera au § 3 du ch. XIV.

9. Comparaison des méthodes et estimation de leur précision. On a étudié deux approches de la résolution des systèmes d'équations linéaires : la méthode de Gauss conduisant à la LU-décomposition et les méthodes des symétries et des rotations fournissant la QR-décomposition. Bien appliquée, la méthode de Gauss exige moins d'opérations arithmétiques et une capacité moindre de la mémoire vive. Elle fournit aussi une solution plus précise au cas d'une bonne sélection de la suite d'éléments principaux ou, ce qui revient au même, d'une heureuse mise à l'échelle de la matrice.

L'unique défaut important de la méthode de Gauss est justement la dépendance de la précision du résultat par rapport à la sélection d'éléments principaux. Autrement dit, les transformations élémentaires utilisées dans la méthode de Gauss modifient la norme de la matrice et il peut s'avérer que les éléments de la matrice $A^{(k)}$, obtenus au cours de la résolution, augmentent fortement en valeur absolue. Cela entraîne l'accroissement du nombre conditionnel et la perte de précision de la solution.

Les méthodes des rotations et des symétries sont assez similaires en précision. Comme on le voit d'après les estimations fournies plus bas, cette précision est à peu près celle du schéma de division unique, bien qu'elle soit moins bonne que la précision du schéma compact à régime d'accumulation. Notons que le régime d'accumulation est aussi nécessaire à la haute précision de la méthode des symétries.

Quant au nombre d'opérations arithmétiques et à la capacité exigée de la mémoire, la méthode des symétries est plus avantageuse que celle des rotations, et par ces caractéristiques ne cède que de peu à la méthode de Gauss.

En même temps, les méthodes des symétries et des rotations sont exemptes du défaut essentiel de la méthode de Gauss, vu que la multiplication par la matrice orthogonale ne change ni la norme spectrale de la matrice A ni son nombre conditionnel spectral.

Les estimations quantitatives des qualités des méthodes étudiées, sur lesquelles s'appuie la comparaison donnée, sont étudiées dans le livre de Voïévodine [40]. Arrêtons-nous sur l'estimation de la précision de la solution.

Trois grandeurs peuvent caractériser la précision de la solution d'un système d'équations linéaires :

a) L'*erreur*, c'est-à-dire $\delta x = x_0 - x$, où x est la solution calculée et x_0 la solution exacte (existant théoriquement). On estime habituellement $\|\delta x\|$ ou l'erreur relative $\|x_0 - x\|/\|x_0\|$.

b) La matrice-colonne $r = b - Ax$ appelée *écart* défini par la solution calculée x. La solution exacte définit un écart nul. L'écart et sa norme $\|b - Ax\|$ peuvent caractériser la précision de la solution. On étudie également l'écart relatif $\|b - Ax\|/\|x\|$.

c) Selon le principe de l'analyse inverse des erreurs d'arrondi, l'influence de ces erreurs sur la solution peut être décrite de la façon suivante. On peut démontrer que la solution approchée est la solution exacte du système d'équations linéaires

$$(A + F)x = b + f.$$

La matrice F s'appelle *perturbation équivalente* de A, et la matrice-colonne f celle de b. Les normes des perturbations équivalentes caractérisent la précision de la solution.

Il est plus commode de calculer l'écart, vu que la solution exacte x_0 nous est inconnue, mais c'est surtout l'erreur qui nous intéresse. L'égalité évidente

$$A\,\delta x = b - Ax = r \tag{14}$$

entraîne

$$\|\delta x\| \leqslant \|A^{-1}\| \cdot \|r\|.$$

Par conséquent, si $\|r\|$ est petite, il en est de même de $\|\delta x\|$. Mais si A^{-1} a une grande norme, il faut exiger un très petit écart pour garantir une faible erreur.

L'approche suivante est possible. Si l'écart r est calculé avec une haute précision (avec utilisation du régime d'accumulation), la solution calculée Δx du système (14) sert de bonne estimation de δx. En outre, $x_1 = x + \Delta x$ est une approximation de la solution exacte meilleure que x. En calculant l'écart défini par x_1, on peut obtenir une approximation suivante x_2, etc. Ce processus sera décrit en détail au paragraphe suivant ; pour le moment, il a fallu s'y référer pour noter que la vitesse de convergence de la suite x, x_1, x_2, ... permet de conclure si la matrice A est plus ou moins bien conditionnée. L'estimation correspondante sera fournie à la p. 433.

Le calcul des approximations successives n'entraîne pas un accroissement sensible du nombre d'opérations et du temps de calcul, car la résolution des systèmes de la forme (14) utilise la LU- ou QR-décomposition déjà obtenue de la matrice. Aussi s'avère-t-il parfois rationnel (voir Wilkinson et Reinsch [43]) de calculer les approximations successives au lieu de chercher le nombre conditionnel et, partant, estimer la précision de la solution.

Passons maintenant à l'estimation de l'erreur relative. Il nous faudra ici se limiter aux formulations.

Si la LU-décomposition est obtenue par la méthode de Gauss, avec choix de l'élément principal suivant la colonne, le produit LU vérifie strictement l'égalité

$$LU = A + G,$$

de sorte que

$$\|G\|_E \leqslant n\rho u \|A\|_E, \tag{15}$$

où u est une erreur d'arrondi relative à une opération, dépendant de la précision du système de calcul utilisé, et ρ se définit ainsi : si $a_{ij}^{(k)}$ sont les éléments des matrices $A^{(k)}$ qui se calculent par élimination, on a

$$\rho = \frac{3}{2}\left(\max_{i,j,k} |a_{ij}^{(k)}|\right)\|A\|_E^{-1}.$$

Ainsi, ρ ne peut être défini que si tous les calculs sont effectués. Sous ce rapport, c'est une estimation *a posteriori*. Elle peut être utilisée comme une estimation *a priori* si l'on tient compte de l'expérience des calculs : si la matrice est bien conditionnée, $\max_{i,j,k} |a_{ij}^{(k)}|$ est seulement de quelques fois plus grand que $\max_{i,j} |a_{ij}|$.

Il s'avère que pour les autres décompositions en facteurs déjà rencontrées on est en mesure de mettre une perturbation équivalente sous la forme analogue à (15) : le produit de matrices obtenu diffère de A par la matrice G dont la norme est majorée par

$$\|G\|_E \leqslant f(n) u \|A\|_E, \tag{16}$$

où la fonction $f(n)$ dépend de l'ordre de la matrice A et de l'algorithme de décomposition utilisé.

Pour la LU-décomposition obtenue d'après le schéma compact de la méthode de Gauss, le terme principal en n de la fonction f est une constante ρ'. Tout comme la constante ρ définie plus haut, ρ' dépend de la croissance des éléments dans les matrices $A^{(k)}$. Cette assertion est vraie à condition que le calcul s'effectue suivant le schéma compact en régime d'accumulation, et indique que dans ce cas on peut s'attendre à un accroissement de précision de n fois par rapport au schéma ordinaire. Remarquons que pour le schéma compact appliqué à une matrice définie positive on a $\rho' = 1$.

L'avantage important des méthodes de recherche de la QR-décomposition réside dans le fait qu'elles permettent d'estimer $f(n)$ *a priori*. Plus précisément, pour la méthode des rotations comme pour la

méthode des symétries, le terme principal de $f(n)$ est de la forme $2{,}9n$. (Pour la méthode des symétries il est obligatoire, comme il a été noté, d'agir en régime d'accumulation.)

On peut montrer que pour toutes les méthodes discutées l'erreur relative de la solution calculée vérifie la majoration

$$\frac{\|x_0 - x\|_E}{\|x_0\|_E} \leqslant 2c_E(A)g(n)u,$$

où les termes principaux en n des fonctions g et f se confondent.

§ 4. Méthodes itératives de résolution des systèmes d'équations linéaires

1. Introduction. Les méthodes directes de résolution des systèmes d'équations linéaires, exposées au § 3 conduisent théoriquement à une solution exacte. On s'occupera maintenant des méthodes *itératives* qui en principe fournissent non pas une solution mais seulement la suite qui converge vers elle. Un terme suffisamment proche de la limite de cette suite est pris pour solution approchée. Ainsi, les méthodes itératives possèdent une erreur théoriquement nécessaire, ce qui ne constitue d'aucune façon pour elles un défaut par rapport aux méthodes directes. En effet, l'erreur mentionnée peut être rendue inférieure à celle due aux erreurs d'arrondi engendrées dans les méthodes directes. Comme on le verra plus bas au point 3, les méthodes itératives peuvent être utilisées pour préciser les solutions obtenues par les méthodes directes.

Le point faible des méthodes itératives réside dans le fait que chacune d'elles ne converge pas, c'est-à-dire ne définit pas une suite convergente, pour tous les systèmes d'équations linéaires. Toutefois, pour une méthode choisie il est souvent possible de transformer le système de manière que cette méthode devienne convergente.

L'efficacité d'une méthode itérative est pour une grande part fonction de sa vitesse de convergence, c'est pourquoi la vitesse de convergence est toujours au centre des préoccupations dans la mise au point de ces méthodes. Il est de même caractéristique que l'efficacité des méthodes itératives est beaucoup plus grande si on commence par une bonne approximation du premier terme de la suite, bien que ces méthodes convergent pour toute approximation initiale.

Il va de soi que la convergence des méthodes itératives n'a lieu que théoriquement. A cause des erreurs d'arrondi inévitables, les approximations calculées diffèrent quelque peu des vraies valeurs. Aussi ne peut-on pas affirmer que les approximations calculées de numéros suffisamment grands se trouvent dans un voisinage aussi petit que l'on veut de la solution

exacte. On peut dire seulement qu'elles appartiennent à son voisinage dont les dimensions sont définies par la précision des calculs. Lorsqu'une approximation tombe dans ce voisinage, les calculs ultérieurs ne peuvent plus augmenter la précision du résultat.

Le domaine type d'application des méthodes itératives sont les systèmes d'équations linéaires apparaissant lors de résolution numérique des équations aux dérivées partielles. Ce sont en général des systèmes qui comportent un grand nombre d'équations à plusieurs inconnues, dont les matrices ne peuvent être écrites et traitées que grâce à leur structure spéciale. Or la structure spéciale des matrices est généralement altérée par les transformations qu'elles subissent dans les méthodes directes. Les méthodes itératives sont exemptes de ce défaut.

Plusieurs méthodes itératives (appliquées non seulement aux systèmes d'équations linéaires) peuvent être obtenues d'après le principe du point fixe (ou principe des applications contractantes). Donnons sa formulation.

Soit $\mathscr{L}$ un espace vectoriel normé complet et soit $\mathbf{F}$ une application (pas forcément linéaire) de $\mathscr{L}$ dans lui-même. On dit que $\mathbf{F}$ est une *application contractante* sur l'ensemble $\mathscr{M} \subseteq \mathscr{L}$ s'il existe un nombre $\alpha \in [0, 1[$ tel que pour tous x' et x'' de $\mathscr{M}$ est satisfaite la condition

$$\|\mathbf{F}(x') - \mathbf{F}(x'')\| \leqslant \alpha \|x' - x''\|.$$

Remarquons que cette définition est applicable aussi lorsque $\mathbf{F}$ est définie non pas sur tout l'espace $\mathscr{L}$ mais seulement sur l'ensemble $\mathscr{M}$.

THÉORÈME 1. *Soit $\mathscr{M}$ un ensemble borné fermé dans $\mathscr{L}$ et soit* $\mathbf{F}$ *une application contractante sur $\mathscr{M}$. Alors l'équation*

$$x = \mathbf{F}(x) \tag{1}$$

admet dans $\mathscr{M}$ une solution et une seule.

Vu qu'on n'aura pas l'occasion d'utiliser cette formulation générale, on n'en donnera pas la démonstration. Le lecteur pourra la trouver, par exemple, dans le livre de Fédoriouk [9]. La démonstration s'appuie sur le fait qu'une fois les conditions du théorème remplies, toute suite de vecteurs $\{x_k\}$ de $\mathscr{M}$, définie par la formule de récurrence

$$x_{k+1} = \mathbf{F}(x_k) \tag{2}$$

converge vers la solution de l'équation (1).

2. Méthode itérative simple. Il est le plus simple de mettre le système d'équations linéaires $Ax = b$ sous la forme (1) par l'addition de x aux deux membres de l'égalité :

$$x = x - Ax + b = (E - A)x + b.$$

On obtient une formule plus générale si au préalable on multiplie les deux

membres de l'égalité par une matrice régulière H :

$$x = x + H(b - Ax),$$

ce qui permet de construire le processus itératif défini par la formule de récurrence

$$x_{k+1} = x_k + H(b - Ax_k). \quad (3)$$

En introduisant le paramètre τ et en désignant H par τB^{-1}, on peut mettre la formule (3) sous la forme

$$B\frac{x_{k+1} - x_k}{\tau} + Ax_k = b. \quad (4)$$

La recherche de la solution approchée avec utilisation de la formule (3) est appelée *méthode itérative simple* ou *méthode stationnaire*. La formule (4) correspond à la *méthode stationnaire implicite générale*. La valeur du paramètre τ est choisie pour le système concret de manière que la vitesse de convergence soit maximale.

On peut donner aux formules (3) (resp. (4)) la forme

$$x_{k+1} = Px_k + f, \quad (5)$$

en posant $P = E - HA$ et $f = Hb$ (resp. $P = E - \tau B^{-1}A$ et $f = \tau B^{-1}b$).

Supposons que le système $Ax = b$ est compatible et que x^* est sa solution. Retranchons x^* des deux membres de l'égalité (5). Compte tenu de ce que $f = Hb = HAx^*$, il vient

$$x_{k+1} - x^* = P(x_k - x^*),$$

d'où en notant $d_k = x_k - x^*$ pour tous les $k = 0, 1, \ldots$, on a

$$d_k = P^k d_0.$$

Il n'est pas difficile de démontrer que la suite $\{d_k\}$ converge pour tout d_0 (et par suite, il en est de même de $\{x_k\}$ pour tout x_0) si et seulement si converge la suite des puissances de la matrice P. En effet, si $P^k \to Q$, il existe pour tout $\varepsilon' > 0$ un numéro k_0 à partir duquel $\|P^k - Q\| < \varepsilon'$, et partant, pour tout d_0

$$\|d_k - Qd_0\| \leqslant \|P^k - Q\|\|d_0\| < \varepsilon'\|d_0\|.$$

Il ne reste qu'à choisir $\varepsilon' = \varepsilon\|d_0\|^{-1}$ pour un $\varepsilon > 0$ arbitraire.

Inversement, si la suite de $P^k d_0$ converge pour tout d_0, en choisissant pour d_0 les colonnes de la matrice unité, on peut montrer que les suites $\{p_i^{(k)}\}$, où $p_i^{(k)}$ est la i-ième colonne de la matrice P^k, convergent pour tous les $i = 1, \ldots, n$. Donc, la suite des puissances converge au sens de la convergence en éléments.

On a vu que $d_k \to Qd_0$, où $Q = \lim_{k\to\infty} P^k$ et, par suite, la limite x de la suite $\{x_k\}$ vérifie la relation

$$x - x^* = Qd_0.$$

D'autre part, il est évident que x est la solution du système et par suite, Qd_0 vérifie le système homogène associé, c'est-à-dire que $AQd_0 = \mathbf{0}$. Cette condition est satisfaite pour tout d_0 si

$$AQ = O,$$

et seulement dans ce cas. On peut maintenant démontrer l'assertion qui suit.

PROPOSITION 1. *Posons que la matrice A du système d'équations linéaires $Ax = b$ est régulière. Dans ce cas, la suite de colonnes définie par la formule de récurrence* (5) *converge pour tout vecteur initial x_0 si et seulement si le rayon spectral de la matrice P est strictement inférieur à l'unité.*

DÉMONSTRATION. Les raisonnements précédents montrent que pour une matrice régulière A la suite $\{x_k\}$ converge pour tout x_0 si et seulement si $Q = O$. Il nous reste à démontrer que $P^k \to O$ si et seulement si le rayon spectral de P est strictement inférieur à l'unité. A cet effet, profitons du théorème 2 du § 3, ch. XII. Appliquée à la fonction $f(\xi) = \xi^k$ et à la matrice P, la formule (15) du § 3, ch. XII, prend la forme

$$P^k = \sum_{i=1}^{s} \sum_{j=1}^{m_i} \frac{\mathrm{d}^{j-1}}{\mathrm{d}\lambda_i^{j-1}} (\lambda_i^k) Z_{ij},$$

où Z_{ij} sont les matrices composantes de la matrice P, s désigne la quantité de nombres caractéristiques distincts, et m_i la multiplicité du i-ième nombre. Il en ressort immédiatement en vertu de l'indépendance linéaire des matrices composantes la proposition nécessaire.

REMARQUE. La démonstration de la proposition 1 montre que le processus itératif est d'autant plus rapide que le rayon spectral de la matrice P est plus petit.

PROPOSITION 2. *Si $\|P\| < 1$, la suite récurrente définie par la formule* (5) *converge pour tout vecteur initial non moins vite que la somme de la progression géométrique de raison $\|P\|$.*

Le fait que la condition $\|P\| < 1$ est suffisante pour la convergence de la suite considérée découle directement de la proposition 1 si l'on se souvient de la majoration du rayon spectral de la matrice (proposition 1, § 4, ch. XII). On donnera une autre démonstration permettant d'estimer la vitesse de convergence. Si l'on désigne pour k quelconque $d_k =$

$= x_k - x_{k-1}$, il est facile de constater que $x_k = x_0 + d_1 + \dots + d_k$. Donc,

$$x = \lim_{k \to \infty} x_k = x_0 + \sum_{k=1}^{\infty} d_k.$$

Supposons que dans $\mathscr{R}_n$ est choisie une norme compatible avec la norme matricielle. En utilisant le critère de Cauchy, il n'est pas difficile de démontrer que, pour qu'une série soit convergente, il suffit que converge la série des normes de ses termes. Pour d_k on a $d_k = Pd_{k-1}$, d'où $\|d_k\| \leqslant \|P\| \cdot \|d_{k-1}\|$ et, dans le cas de $\|P\| < 1$, pour la série des normes est vérifié le critère de d'Alembert. Ainsi donc, la vitesse de convergence de la suite $\{x_k\}$ est dans ce cas supérieure ou égale à celle de la somme de la progression géométrique de raison $\|P\|$. La proposition est démontrée.

On a parlé plus haut du rayon spectral de la matrice comme d'une grandeur bien déterminée. En réalité, si les opérations arithmétiques s'effectuent avec arrondissement, le rayon spectral, comme nombre d'autres grandeurs, ne peut être déterminé de façon précise. En effet, soit un nombre λ tel que la matrice $P - \lambda E$ est quasi singulière. Le nombre λ se comporte dans tous les calculs de la même façon que le nombre caractéristique. En particulier, si $\lambda > 1$, la méthode itérative simple de matrice P ne converge pas avec la précision exigée.

Dans tous les cas, pour accélérer la convergence, le processus doit être reglé de manière que la norme de la matrice P soit la plus petite possible. Si l'on revient aux formules (3) (resp. (4)), cela signifie qu'il faut choisir la matrice H (resp. le paramètre τ et la matrice B) de manière que la norme $\|E - HA\|$ (resp. $\|E - B^{-1}\tau A\|$) soit la plus petite possible.

Si on avait pu poser $H = A^{-1}$, le processus (à condition d'exécution précise des opérations) aurait convergé en une itération. En utilisant les propriétés de la matrice donnée A, on choisit, dans les différentes méthodes itératives, les matrices qui, d'une façon ou autre, se rapprochent de A^{-1}.

La *méthode de Jacobi* bien connue se rapporte aux systèmes dont les matrices possèdent une diagonale principale dominante (voir § 4, ch. XII). Notons D la matrice diag $(a_{11}, \dots, a_{nn})$, c'est-à-dire la matrice diagonale dont les éléments diagonaux sont égaux aux éléments homologues de la matrice A. La méthode de Jacobi est définie par la formule (3) où $H = D^{-1}$. Selon la proposition 1, la suite construite d'après cette méthode converge si et seulement si le rayon spectral de la matrice $E - D^{-1}A$ est strictement inférieur à l'unité.

Montrons que pour la convergence de la méthode de Jacobi il suffit que A possède une diagonale principale dominante. Pour cela, estimons la c-norme de la matrice $E - D^{-1}A$. Cette matrice a sur la diagonale princi-

pale des zéros, et ses éléments non diagonaux sont $\alpha_{ik} = -\frac{a_{ik}}{a_{ii}}$. Donc on a

$$\| E - D^{-1}A \|_c = \max_i \sum_k |\alpha_{ik}| = \max_i \sum_{k \neq i} \left| \frac{a_{ik}}{a_{ii}} \right|.$$

Si A a une diagonale principale dominante, chacune de ces sommes est strictement inférieure à l'unité, de sorte que $\| E - D^{-1}A \|_c < 1$.

L'utilisation d'autres normes fournira des conditions suffisantes différentes.

Une série de méthodes itératives simples spéciales est liée à des matrices symétriques définies positives. Pour toute matrice A de ce genre on peut toujours trouver une matrice H pour que le processus itératif (3) converge. En effet, les nombres caractéristique de A sont compris dans un intervalle $]0, \alpha[$. En posant $H = \frac{2}{\alpha} E$, on obtient la matrice $P = E - \frac{2}{\alpha} A$, dont les nombres caractéristiques appartiennent à l'intervalle $]-1, 1[$. En effet, si $A = S^{-1}A'S$, où $A' = \text{diag}(\lambda_1, \ldots, \lambda_n)$, on a $P = S^{-1}P'S$, où P' est la matrice diagonale avec éléments de la forme $1 - \frac{2}{\alpha}\lambda_i$ sur la diagonale principale.

On n'exposera pas d'une façon détaillée l'application de la méthode itérative simple et de la méthode stationnaire implicite générale aux matrices définies positives. Cet exposé peut être trouvé dans les livres de Faddeev et Faddeeva [8] et de Samarski [34].

Notons que le cas d'une matrice symétrique définie positive est assez général, car le système compatible $Ax = b$ est équivalent au système ${}^tAAx = {}^tAb$ à matrice symétrique définie positive. Toutefois, on a vu à la p. 392 que la matrice ${}^tAA$ est en général moins bien conditionnée que la matrice A. Aussi une telle transformation du système, effectuée pour pouvoir appliquer les méthodes itératives, n'est-elle pas d'une grande importance pratique. L'intérêt suscité par l'utilisation des méthodes itératives dans la résolution des systèmes à matrices définies positives s'explique par le fait que ces systèmes sont assez souvent rencontrés dans les applications.

3. Précision itérative. Supposons qu'en utilisant la méthode de Gauss on ait obtenu la LU-décomposition de la matrice A. Les erreurs d'arrondi ont altéré le résultat et, par suite, $LU \neq A$ pour les matrices calculées L et U. Dans la méthode stationnaire implicite générale on peut poser $B = LU$ et $\tau = 1$. La formule (4) prendra alors la forme

$$LUd_{k+1} = r_k, \tag{6}$$

où $r_k = b - Ax_k$ est le k-ième écart et $d_{k+1} = x_{k+1} - x_k$, la

$(k + 1)$-ième correction. Pour approximation initiale on choisit la solution du système $LUx_0 = b$ et on calcule l'écart correspondant r_0. On le prend pour colonne des termes constants du système ayant la même matrice LU. La solution de ce système est la première correction d_1. L'approximation suivante x_1 s'obtient comme $x_0 + d_1$. Ensuite, on recherche l'écart de cette solution $r_1 = b - Ax_1$ et la deuxième correction d_2 comme solution du système $LUd_2 = r_1$. La deuxième approximation sera $x_2 = x_1 + d_2$, etc.

Si le produit LU est proche de A, la norme $\| E - (LU)^{-1}A \|$ est petite, de sorte que le processus converge très vite à condition que les calculs soient effectués de façon exacte. S'ils ont la même précision que la LU-décomposition, ils ne donnent aucune précision, car dans ce cas toutes les approximations sont aggravées d'une erreur du même ordre que l'erreur de l'approximation initiale.

Pratiquement, cela signifie que si la précision est réalisée par itération, les calculs doivent s'effectuer avec une précision double. Comme le montre une analyse plus détaillée (voir Forsythe et Moler [10]), la précision est particulièrement nécessaire dans le calcul des écarts et dans l'addition des corrections. Si ces calculs s'effectuent à $2q$ chiffres près (ce qui peut être réalisé avec un régime d'accumulation) et les autres calculs à q chiffres près, la précision itérative permet d'obtenir une solution exacte, arrondie jusqu'à q chiffres.

Montrons que la convergence de la précision itérative est fonction du conditionnement de la matrice A et de la précision de la décomposition obtenue. Supposons que le produit des matrices calculées L et U vérifie l'égalité $LU = A + G$. Alors,

$$(LU)^{-1}A = (A + G)^{-1}A = (E + A^{-1}G)^{-1}.$$

En posant que $\| A^{-1} \| \cdot \| G \| < 1$, on peut représenter cette matrice sous la forme d'une série

$$E - A^{-1}G + (A^{-1}G)^2 - \ldots$$

Pour la matrice $E - (LU)^{-1}A$ dont dépend la convergence, on obtient la série

$$A^{-1}G - (A^{-1}G)^2 + \ldots$$

Chaque somme partielle de cette série ne dépasse pas en norme la somme partielle correspondante de la série des normes, et l'on obtient

$$\| E - (LU)^{-1}A \| \leqslant \frac{\| A^{-1} \| \cdot \| G \|}{1 - \| A^{-1} \| \cdot \| G \|}.$$

Selon la formule (15) du § 3, $\| G \|_E \leqslant \theta \| A \|_E$, où $\theta = n\rho u$. Donc, pour la

convergence du processus de précision itérative, il suffit que

$$\frac{\theta c_E(A)}{1 - \theta c_E(A)} < 1.$$

On a déjà supposé que $\|A^{-1}\| \cdot \|G\| < 1$, c'est-à-dire que $\theta c_E(A) < 1$. Avec cette hypothèse, la condition précédente signifie que $\theta c_E(A) < \frac{1}{2}$. Ainsi donc, on a démontré la

PROPOSITION 3. *Le processus de précision itérative converge si* (*dans les notations du point* 9, § 3)

$$2n\rho u c_E(A) < 1.$$

La précision itérative est assurément possible si au lieu de la LU-décomposition, est donnée une autre, par exemple la QR-décomposition.

4. Méthode de Seidel. Dans la méthode stationnaire implicite générale il est important que la matrice B soit facilement inversible. Choisissons donc une matrice triangulaire. Si les éléments diagonaux de la matrice A sont différents de zéro, la matrice triangulaire inférieure

$$\tilde{L} = \begin{Vmatrix} a_{11} & 0 & \dots & 0 \\ a_{21} & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{Vmatrix}$$

(déduite de A par substitution des zéros aux éléments situés au-dessus de la diagonale) est régulière. La méthode de Seidel consiste dans la mise en œuvre de la formule (4) pour $\tau = 1$ et $B = \tilde{L}$, c'est-à-dire de

$$\tilde{L}(x_{k+1} - x_k) + Ax_k = b,$$

ou

$$\tilde{L}x_{k+1} + Ux_k = b, \tag{7}$$

où la matrice $U = A - \tilde{L}$ diffère de A par le fait que les éléments situés sur et sous la diagonale sont remplacés par des zéros. La formule (7) permet de calculer facilement x_{k+1} d'après x_k.

Selon la proposition 1, la méthode de Seidel converge si et seulement si tous les nombres caractéristiques de la matrice $E - \tilde{L}^{-1}A = -\tilde{L}^{-1}U$ sont en module inférieurs à l'unité. On vérifie aisément que ces nombres se confondent avec les racines de l'équation

$$\det(U + \lambda\tilde{L}) = 0. \tag{8}$$

PROPOSITION 4. *Pour que la méthode de Seidel converge il suffit que la matrice A soit symétrique et définie positive.*

Cette proposition est un cas particulier de la proposition 6 qu'on démontrera plus loin.

La méthode de Seidel converge plus vite que la méthode de Jacobi et exige une mémoire de moindre capacité. Cette dernière circonstance est liée au fait qu'avec l'utilisation des matrices triangulaires le calcul de la i-ième composante du vecteur x_{k+1} rend inutile la composante homologue du vecteur x_k (voir formule (7)).

Au point suivant, on étudiera un procédé qui fait accélérer la convergence de la méthode de Seidel par introduction d'un paramètre.

5. **Méthode de relaxation supérieure.** Subdivisons la matrice A en trois matrices $A = L + D + U$. La matrice D est comme plus haut égale à $\operatorname{diag}(a_{11}, \ldots, a_{nn})$ et L et U se déduisent de A par substitution des zéros aux éléments a_{ik} pour $i \geqslant k$ et $i \leqslant k$ respectivement. Ainsi, la matrice $\tilde{L}$ du point précédent vaut $L + D$. Supposons que $\det D \neq 0$. La méthode de relaxation supérieure se définit par la formule (4) à condition que $\tau = \omega$ et $B = D + \omega L$. La formule (4) prend la forme

$$(D + \omega L)(x_{k+1} - x_k) + \omega A x_k = \omega b,$$

ou

$$(D + \omega L)x_{k+1} = [(1 - \omega)D - \omega U]x_k + \omega b. \tag{9}$$

Pour $\omega = 1$ on en tire la formule (7) traduisant la méthode de Seidel.

En appliquant la proposition 1, on voit que la méthode de relaxation supérieure converge si et seulement si le rayon spectral de la matrice

$$P = E - (D + \omega L)^{-1}\omega A$$

est strictement inférieur à l'unité. La matrice P peut prendre une autre forme

$$(D + \omega L)^{-1}[(1 - \omega)D - \omega U].$$

L'équation caractéristique de la matrice P peut être transformée de la façon suivante :

$$\det(P - \lambda E) = \det(D + \omega L)^{-1} \det[(1 - \omega)D - \omega U - \lambda(D + \omega L)].$$

D'où l'on voit que les nombres caractéristiques de la matrice P se confondent avec les racines de l'équation

$$\det[(1 - \omega)D - \omega U - \lambda(D + \omega L)] = 0. \tag{10}$$

Portons dans l'équation (10), au lieu de la variable λ, la variable ξ liée à

λ par la relation

$$\lambda = \frac{\xi + 1}{\xi - 1}.$$

On voit aussitôt que la condition $|\lambda| < 1$ est équivalente à la condition $\mathrm{Re}\, \xi < 0$ et que l'équation (10) se met sous la forme

$$\det\, [-\omega A \xi - (2 - \omega) D + \omega(U - L)] = 0. \tag{11}$$

On a ainsi obtenu la

PROPOSITION 5. *La méthode de relaxation supérieure converge si et seulement si les parties réelles de toutes les racines de l'équation* (11) *sont strictement négatives.*

Cela permet d'obtenir la condition suffisante suivante.

PROPOSITION 6. *Pour que la méthode de relaxation supérieure soit convergente il suffit que la matrice A soit symétrique et définie positive et que* $\omega \in]0, 2[$.

Pour le démontrer, écrivons l'équation (11) pour une matrice symétrique A et considérons la racine ξ_0 de cette équation. Il existe pour ξ_0 une matrice-colonne non nulle (en général, complexe) x telle que

$$[-\omega A \xi_0 - (2 - \omega) D + \omega(U - L)]x = \mathbf{0},$$

et par suite,

$$-\omega \xi_0 {}^t\bar{x} A x - (2 - \omega) {}^t\bar{x} D x + \omega {}^t\bar{x}(U - L)x = 0. \tag{12}$$

La matrice $U - L$ est symétrique gauche et l'on a pour elle

$${}^t\bar{x}(U - L)x = {}^t(-{}^t\bar{x}(U - L)x) = -{}^t x(U - L)\bar{x} = -{}^t x(U - L)x.$$

Aussi le dernier terme dans l'égalité (12) possède-t-il la partie réelle nulle. On démontre de façon analogue que ${}^t\bar{x}Ax$ et ${}^t\bar{x}Dx$ sont réelles. En outre, si A est définie positive, tous ses éléments diagonaux sont strictement positifs, et la forme quadratique ${}^t\bar{x}Dx$ est aussi définie positive. En annulant la partie réelle de (12), on obtient

$$\mathrm{Re}\, \xi_0 = \frac{\omega - 2}{\omega} \frac{{}^t\bar{x}Dx}{{}^t\bar{x}Ax}. \tag{13}$$

Si $\omega \in]0, 2[$, on a $(\omega - 2)/\omega < 0$. Il en découle que $\mathrm{Re}\, \xi_0 < 0$. La proposition est démontrée.

La proposition 4 s'ensuit comme un cas particulier pour $\omega = 1$.

La formule (12) permet d'exprimer quelques considérations sur le choix du paramètre ω dans le cas d'une matrice A symétrique définie positive. Supposons que les racines de (11) sont réelles et mettons (13) sous la forme $\xi_0 = -\alpha\rho$, où $\alpha = (2 - \omega)/\omega$ et $\rho = ({}^t\bar{x}Dx)/({}^t\bar{x}Ax)$. A la valeur de

$\lambda = 0$ correspond la valeur de $\xi = -1$. On doit choisir α de manière que ξ soit le plus rapproché possible de -1.

Estimons ρ. On peut à cet effet se servir du résultat obtenu dans le § 2 du ch. XI, d'après lequel ρ est compris entre les racines minimale et maximale de l'équation $\det(D - \lambda A) = 0$. Notons ses racines respectivement ρ_1 et ρ_2. Alors toutes les racines de l'équation (11) sont comprises entre $-\alpha\rho_2$ et $-\alpha\rho_1$. Pour rapprocher ces bornes de -1, on doit choisir α de manière que le plus grand des nombres $|1 - \alpha\rho_2|$ et $|1 - \alpha\rho_1|$ soit minimal. Ceci est équivalent à l'exigence que soit minimal le plus grand des nombres $|\alpha - \rho_1^{-1}|$ et $|\alpha - \rho_2^{-1}|$. Cette dernière condition sera remplie si α est le milieu du segment $[\rho_2^{-1}, \rho_1^{-1}]$. Ce choix de α n'est pas le meilleur mais il fournit des résultats acceptables. Le problème de la recherche du meilleur α est résolu sous forme générale mais cette résolution est assez compliquée et on n'y s'arrêtera pas.

§ 5. Calcul des vecteurs propres et des valeurs propres

1. Remarques préliminaires. Dans le présent paragraphe on étudiera le problème de recherche des solutions non nulles du système d'équations

$$Ax = \lambda x,$$

qu'on appellera par abus de langage *vecteurs propres de la matrice A*. Le problème peut être posé de deux façons différentes la recherche de tous les nombres caractéristiques et des vecteurs propres associés, ce qu'on appelle *problème complet des valeurs propres,* et la recherche d'un seul ou de quelques-uns des ces derniers, appelée *problème partiel des valeurs propres.*

Il va de soi que le problème des valeurs propres, aussi bien complet que partiel, est beaucoup plus compliqué que le problème de résolution du système d'équations linéaires, étudié plus haut. Parmi le grand nombre d'algorithmes proposés pour sa résolution, il n'en existe aucun qu'on puisse recommander dans tous les cas, bien que certains d'entre eux peuvent être dégagés parmi les autres par leur plus grande efficacité. Quelques algorithmes sont particulièrement efficaces dans des cas spéciaux, par exemple, sont spécialement adaptés aux matrices symétriques ou en bandes.

Toutes les méthodes existantes de résolution du problème des valeurs propres peuvent être divisées en deux grands groupes : les méthodes directes qui se basent sur la résolution de l'equation caractéristique et les méthodes itératives. Dans les méthodes directes, une étape importante est la recherche des coefficients du polynôme caractéristique, vu que leur calcul d'après les formules générales, directement déduites de la définition, exige

un très grand nombre d'opérations arithmétiques. Les calculs nécessaires à la recherche du polynôme caractéristique sont d'habitude utilisés pour obtenir les vecteurs propres. Le résultat obtenu par les méthodes directes est en principe approché (même si on néglige les erreurs d'arrondi) car les racines d'un polynôme caractéristique ne peuvent être trouvées que d'une façon approchée. On n'exposera pas ici les méthodes directes de résolution du problème des valeurs propres. Elles sont exposés dans les livres de Gantmacher [12], de Faddeev et Faddeva [8].

La méthode itérative typique est la *méthode des rotations,* ou *méthode de Jacobi.* La matrice symétrique initiale A est soumise à une série de transformations de rotation, étudiées au § 3 avec la construction de la QR-décomposition. A la différence de la QR-décomposition, la transformation s'effectue à chaque étape suivant la formule

$$A^{(k+1)} = T_{k+1}^{-1} A^{(k)} T_{k+1},$$

où $A^{(k)}$ est une matrice obtenue après la k-ième étape, et T_{k+1} une matrice de rotation. Cette formule ne permet pas en un nombre fini d'étapes de réduire la matrice à la forme pour laquelle les nombres caractéristiques puissent s'établir directement. La $(k + 1)$-ième rotation est choisie de façon à annuler le plus grand en module élément non diagonal de la matrice $A^{(k)}$, ou si c'est difficile, tout élément supérieur à un certain nombre donné. En d'autres variantes, on diminue la somme des carrés d'éléments non diagonaux ou, en général, une norme quelconque de la matrice $\hat{A}^{(k)}$ obtenue à partir de $A^{(k)}$ par substitution des zéros aux éléments diagonaux.

En fin de compte, après un nombre suffisant s d'étapes, la matrice $A^{(s)}$ s'avère proche de la matrice diagonale et possède les mêmes nombres caractéristiques que A. Les éléments diagonaux de $A^{(s)}$ sont pris pour des valeurs approchées des nombres caractéristiques et les colonnes de la matrice $T = T_1 \ldots T_s$, pour des vecteurs propres approchés.

On ne s'arrêtera pas davantage sur la méthode de Jacobi. Remarquons qu'à cause de l'énorme quantité de travail qu'elles exigent, cette méthode ainsi que les autres méthodes itératives n'ont subi une grande extension qu'après l'apparition des ordinateurs. C'est une des méthodes les plus efficaces pour des matrices symétriques.

2. Méthode des puissances. Cette méthode s'applique à la résolution du problème partiel des valeurs propres. Dans la plus simple des variétés (*méthode des puissances directe sans décalages*) on construit à partir d'un vecteur arbitraire x_0 une suite de vecteurs $x_0, x_1, \ldots$ tels que

$$Ax_k = y_{k+1}, \quad x_{k+1} = \alpha_{k+1}^{-1} y_{k+1}, \tag{1}$$

le facteur strictement positif α_{k+1} étant choisi de manière que $\|x_{k+1}\| = 1$.

Il va de soi que la même suite peut être définie par la formule

$$\beta_k x_k = A^k x_0, \tag{2}$$

où $\beta_k = \alpha_1 \dots \alpha_k$.

Le choix entre les expressions (1) on (2) est dicté par l'économie du nombre d'opérations arithmétiques. Généralement, on utilise la formule (1) car pour obtenir, disons, x_2 par la formule (2) il faut multiplier $n + 1$ fois la matrice A par la matrice-colonne, tandis qu'avec la formule (1), 2 fois suffisent. Toutefois, pour de grandes puissances, la calcul successif de $A^2, A^4, A^8, \dots$ permet d'obtenir le résultat par la formule (2) en un moins grand nombre d'opérations.

Supposons que parmi les nombres caractéristiques de la matrice A il existe un nombre réel λ_1 dépassant en module tous les autres. A cette condition, on peut démontrer que la suite construite est convergente. Ecrivons la décomposition spectrale de A^k suivant la formule (15) du § 3, ch. XII. On obtient

$$\beta_k x_k = \sum_{i=1}^{s} \sum_{j=1}^{m_i} (\lambda_i^k)^{(j-1)} Z_{ij} x_0, \tag{3}$$

où m_i sont les multiplicités des racines λ_i du polynôme minimal et Z_{ij} les matrices constantes appelées matrices composantes. Divisons les deux membres de l'égalité par $k^{s_1 - 1}\lambda_1^k$. Il vient alors

$$\gamma_k x_k = \sum_{i=1}^{m} \sum_{j=1}^{s_i} \nu_{kij} Z_{ij} x_0, \tag{4}$$

où $\gamma_k = \beta_k k^{-s_1+1} \lambda_1^{-k}$. En effectuant la dérivation, on peut mettre les coefficients ν_{kij} sous la forme suivante :

$$\nu_{kij} = \frac{k(k-1)\dots(k-j+2)}{k^{s_1-1}\lambda_1^{j-1}} \left(\frac{\lambda_i}{\lambda_1}\right)^{k-j+1}.$$

Pour $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$, il est évident que $\nu_{kij} \to 0$ pour $k \to \infty$. Pour $i = 1$, avec $j < s_1$, on a de même $\nu_{k1j} \to 0$, mais cette fois parce que la puissance de k dans le numérateur est strictment inférieure à $s_1 - 1$. Le coefficient ν_{k1s_1} a une limite finie non nulle qu'on notera μ. Donc, pour $k \to \infty$

$$\sum_{i=1}^{m} \sum_{j=1}^{s_i} \nu_{kij} Z_{ij} x_0 \to \mu Z_{1s_1} x_0.$$

On voit que pour tour $\varepsilon > 0$ et tout $k > k_0(\varepsilon)$, on a

$$\|\gamma_k x_k - \mu Z_{1s_1} x_0\| < \varepsilon,$$

d'où on obtinet sans difficulté

$$\|\mu Z_{1s_1} x_0\| - \varepsilon < \|\gamma_k x_k\| < \|\mu Z_{1s_1} x_0\| + \varepsilon$$

et, comme $\|\gamma_k x_{k+1}\| = \gamma_k \|x_{k+1}\| = \gamma_k$,

$$\gamma_k \to \|\mu Z_{1s_1} x_0\|.$$

Dans le cas général, $Z_{1s_1} x_0 \neq 0$, et par suite, la limite γ de la suite γ_k n'est pas égale à zéro. D'où

$$x_{k+1} \to \frac{\mu}{\gamma} Z_{1s_1} x_0.$$

A l'étude du cas $Z_{1s_1} x_0 = 0$ on reviendra un peu plus tard.

Pour une suite convergente $\{\gamma_k\}$, le rapport γ_k/γ_{k-1} tend vers 1, de sorte que

$$\frac{\gamma_k}{\gamma_{k-1}} = \frac{\alpha_k}{\lambda_1} \frac{k-1}{k}$$

entraîne que $\alpha_k \to \lambda_1$. Les égalités (1) peuvent être écrites sous la forme

$$A x_k = \alpha_{k+1} x_{k+1}.$$

En passant à la limite on voit que x_k tend vers le vecteur propre associé à la valeur propre λ_1.

D'ailleurs, en se souvenant de la définition des matrices composantes, on remarque immédiatemant que dans le cas où $Z_{1s_1} x_0$ est différent de zéro, il est le vecteur propre associé à λ_1.

Considérons de façon plus détaillée le raisonnement qu'on vient de faire. Chaque terme de la somme (3) tend vers 0 ou vers ∞ pour $k \to \infty$, ce qui est dû à son « ordre » $k^{j-1}\lambda_i^k$. On a imposé la restriction $|\lambda_1| > |\lambda_i|$ $(i > 1)$. Dans le cas général, il peut arriver que plusieurs termes de la somme aient le même ordre. La suite de x_k converge alors vers une combinaison linéaire de vecteurs propres associés à ces termes.

Pour les matrices réelles, l'hypothèse que le nombre caractéristique module maximal soit réel est très importante. La méthode des puissances permet de trouver les nombres caractéristiques complexes et les sous-espaces invariants correspondants pour les matrices réelles, mais on ne s'arrêtera pas sur cette question.

Discutons maintenant le cas où le vecteur initial x_0 est choisi de manière que $Z_{1s_1} x_0 = 0$. Dans ce cas, le terme de la somme (3) dont l'ordre est maximal s'annule. En modifiant le raisonnement, on devra diviser les deux membres de l'égalité (3) par l'ordre maximal des termes non nuls. Pour le reste la démonstration sera la même et on verra que la suite de x_k tend vers le vecteur propre correspondant au terme dominant. En particulier, il peut arriver que ce vecteur propre soit associé à une autre valeur propre.

Cependant en réalité, l'égalité $Z_{1s_1}x_0 = \mathbf{0}$ ne peut être vérifiée exactement. Même si elle se vérifiait, les erreurs d'arrondi aux premières itérations engendreraient un effet équivalent à sa perturbation. Par suite, dans le cas considéré, la suite de x_k peut assurément converger vers le vecteur propre associé au nombre caractéristique de module maximal. Il est possible que cela ne se produise pas si les erreurs d'arrondi sont petites et la suite converge rapidement.

3. Méthode des puissances inverse. Les autres variantes de la méthode des puissances utilisent le même processus qui s'applique non pas à la matrice initiale A mais à une fonction de cette matrice. Considérons une variante largement appliquée qu'on appelle *méthode des puissances inverse avec décalage.* Dans ce cas, à la place de la matrice A on utilise la matrice $(A - \rho E)^{-1}$. On peut évidemment ne pas calculer la matrice inverse et construire la suite d'après les formules

$$x_k = \alpha_{k+1}(A - \rho E)x_{k+1}, \tag{5}$$

en résolvant le système d'équations linéaires en x_{k+1}. Ceci étant, on se sert de la LU-décomposition ou de la QR-décomposition de la matrice $A - \rho E$, calculée une fois.

Les nombres caractéristiques μ_i de la matrice $A - \rho E$ sont liés aux nombres caractéristiques de A par les égalités $\mu_i = (\lambda_i - \rho)^{-1}$. Ainsi, la méthode des puissances inverse avec décalage fournit une suite convergeant vers le vecteur propre de A qui est associé à celui des nombres caractéristiques λ_l pour lequel $|\lambda_l - \rho|^{-1}$ est maximal. Ce nombre caractéristique est le plus proche du nombre ρ.

On utilise en général la méthode des puissances inverse avec décalage quand il faut calculer le vecteur propre d'après la valeur approchée du nombre caractéristique. On admet dans ce cas que le décalage est égal à cette valeur, et la suite converge très rapidement, peut-être en une seule itération. La difficulté réside ici dans le fait que pour un ρ proche du nombre caractéristique λ_l on est obligé de résoudre le système dont la matrice est mal conditionnée.

Montrons que malgré cette difficulté, le vecteur propre peut être obtenu de façon assez exacte s'il n'est pas mal conditionné, et que la valeur propre λ_l est bien séparée des valeurs voisines.

La précision du vecteur calculé x peut être estimée d'après l'écart

$$v = (A - \rho E)x.$$

La norme $\|x\| = ({}'xx)^{1/2}$ du vecteur x sera considérée égale à 1. Alors, l'égalité précédente peut être mise sous la forme

$$(A - v'x)x = \rho x,$$

d'où l'on voit que x est un vecteur propre exact associé à la valeur propre ρ de la matrice perturbée $A - v'x$. Si la norme de l'écart est petite, il en est de même de celle de la perturbation $H = -v'x$. En effet, pour tout vecteur y tel que $\|y\| = 1$ on a $\|(v'x)y\| = \|v('xy)\| \leqslant \|v\|$, l'égalité étant réalisée pour $x = y$. On trouve donc pour la perturbation H

$$\|H\| = \sup_{|y|=1} \|Hy\| = \|v\|.$$

La formule (5) peut être écrite sous la forme

$$(A - \rho E)x_{k+1} = \frac{x_k}{\alpha_{k+1}}.$$

Rappelons que α_{k+1} est choisi de manière que le vecteur x_{k+1} ait la norme égale à 1, c'est-à-dire telle que $\alpha_{k+1} = \|(A - \rho E)^{-1}x_k\|$. Le vecteur x_k/α_{k+1} peut être interprété comme l'écart engendré par x_{k+1} et par suite, plus α_{k+1} est grand, et plus l'écart est petit et moins la perturbation équivalente H est grande.

Pour estimer la grandeur α_k, rappelons-nous que dans le cas général (c'est-à-dire pour le vecteur initial x_0 ne vérifiant pas une égalité exacte) α_k tend vers le plus grand en module nombre caractéristique de la matrice définissant le processus. Si le décalage ρ est proche de λ_l et est éloigné des autres nombres caractéristiques de A, le nombre caractéristique de plus grand module de la matrice $(A - \rho E)^{-1}$ est $(\lambda_l - \rho)^{-1}$. Ainsi,

$$\alpha_k = (\lambda_{l-\rho})^{-1}O(1),$$

et l'on peut s'attendre à de grandes valeurs de α_k et, partant, à une petite norme de l'écart.

Cependant aux premières itérations, cette relation limite n'est pas aussi importante que le choix du vecteur initial. Montrons que dans le cas d'un heureux choix de ce vecteur on peut obtenir un petit écart déjà à la première itération. En effet, on peut admettre que la norme de la matrice $A - \rho E$ n'est pas grande et que le nombre conditionnel $\|A - \rho E\| \cdot \|(A - \rho E)^{-1}\|$ est élevé car $\|(A - \rho E)^{-1}\| > \varepsilon^{-1}$, où ε est un petit nombre. Dans le cas de la norme spectrale cela signifie qu'il existe un vecteur unitaire x_0 pour lequel $|(A - \rho E)^{-1}x_0| > \varepsilon^{-1}$. C'est le premier vecteur de la base singulière de la matrice $(A - \rho E)^{-1}$. Si l'on prend x_0 pour vecteur initial, la norme de l'écart devient strictement déjà après la première étape.

Le raisonnement fait est peu réconfortant vu qu'un bon vecteur initial nous est inconnu. Mais il n'est pas si inutile que cela peut paraître. En effet, si le système d'équations linéaires est mal conditionné, c'est la composante de l'erreur en premier vecteur de la base singulière de la matrice

inverse qui s'avère la plus grande (proposition 5, § 2). Par conséquent, si à la première itération la solution comporte une grande erreur, celle-ci entraîne une diminution de l'ecart à l'itération suivante.

Discutons la vitesse de convergence de la méthode des puissances inverse. Elle est fonction des coefficients dans la décomposition (3) écrite pour la matrice $(A - \rho E)^{-1}$:

$$\beta_k x_k = \sum_{i=1}^{s} \sum_{j=1}^{m_i} [(\lambda_i - \rho)^{-k}]^{(j-1)} W_{ij} x_0.$$

La dérivée $[(\lambda_i - \rho)^{-k}]^{(j-1)}$ est égale à $(\lambda_i - \rho)^{-k}$ pour $j = 1$, et à

$$\frac{(-1)^k k(k+1) \dots (k+j-2)}{(\lambda_i - \rho)^{k+j-1}}$$

pour $j > 1$. Si ρ est proche de λ_l et que $W_{ls_l} x_0 \neq \mathbf{0}$, le terme contenant ce vecteur est dominant et l'ordre de son coefficient est $\dfrac{k^{s_l-1}}{(\lambda_l - \rho)^{k+s_l-1}}$. La vitesse de convergence se définit par la vitesse à laquelle le rapport

$$\frac{k^{j-1}(\lambda_l - \rho)^{k+s_l-1}}{k^{s_l-1}(\lambda_i - \rho)^{k+j-1}}$$

tend vers zéro. On voit que le cas le moins favorable est celui d'existence du nombre caractéristique $\lambda_{l'}$, également proche de ρ et de grande multiplicité dans le polynôme minimal. Parfois on peut corriger la situation en variant le décalage, mais si λ_l et $\lambda_{l'}$ sont très proches l'un de l'autre, cela permet de dire que le problème est mal conditionné.

La vitesse de convergence diminue de même si le vecteur $W_{ls_l} x_0$ est petit ou égal à zéro. Comme il a été montré plus haut, son annulation n'influe en général pas sur le vecteur propre calculé.

Le vecteur vers lequel converge la suite peut dépendre de l'approximation initiale dans deux cas. En premier lieu, s'il existe plusieurs nombres caractéristiques différents dont les modules sont égaux et dépassent les modules des autres nombres. En second lieu, si au nombre caractéristique maximal en module est associé un sous-espace propre de dimension $m > 1$. On décrira un peu plus loin comment obtenir dans ce cas la base de ce sous-espace.

Les erreurs d'arrondi apparaissant à chaque itération comme dans le cas de résolution itérative du système d'équations linéaires influent sur la convergence du processus.

Si l'on tient compte des erreurs d'arrondi et que dans l'égalité

$(A - \rho E)^{-1}x_k = \alpha_{k+1}x_{k+1}$ on comprenne par x_{k+1} et x_k des vecteurs en fait calculés, cette égalité doit être corrigée par un terme complémentaire, soit :

$$(A - \rho E)^{-1}x_k = \alpha_{k+1}x_{k+1} + u_k.$$

Tous les vecteurs u_k sont bornés : $\| u_k \| < \delta$, avec δ indépendant de l'itération. Comme plus haut, on peut écrire

$$((A - \rho E)^{-1} - u_k{}'x_k)x_k = \alpha_{k+1}x_{k+1}.$$

Cela signifie que le résultat est le même que dans l'itération exacte avec matrice perturbée $(A - \rho E)^{-1} + F$, où $F = -u_k{}'x_k$. Aussi, à une des itérations, peut-il s'avérer que le vecteur x_{k+1} coïncide avec x_k en précision d'exécution, mais on ne pourra pas affirmer qu'il est un vecteur propre de $(A - \rho E)^{-1}$ (et, par suite, de A), il le sera pour une autre matrice voisine. Autrement dit, on ne peut pas garantir que l'amélioration des approximations x_k se poursuivra après que x_k devienne égal au vecteur propre d'une matrice $(A - \rho E)^{-1} = H$, où $\| H \| < \delta$. La précision réellement accessible dépend du conditionnement du vecteur propre et de la valeur propre.

Si la valeur approchée du nombre caractéristique est inconnue *a priori*, la méthode des puissances convergera en principe pour tout décalage (par exemple, nul), mais cette convergence sera beaucoup moins rapide. Dans ce cas, pour accélérer la convergence, on recourt au décalage qui varie à chaque itération. Pour une matrice symétrique, les meilleurs résultats s'obtiennent si pour décalage on prend le quotient de Rayleigh calculé pour chaque vecteur x_k. On ne s'arrêtera pas en détail sur ce sujet, comme sur de nombreux autres procédés d'accélération et de précision de la méthode des puissances. Notons seulement qu'avec toutes ces améliorations la méthode des puissances inverse avec décalage s'est avérée la plus précise et efficace pour obtenir un ou plusieurs vecteurs propres.

4. Développement ultérieur de la méthode des puissances. La méthode des puissances décrite plus haut permet d'obtenir un des vecteurs propres associés au nombre caractéristique maximal en module. Voyons comment elle peut être modifiée afin d'obtenir les autres vecteurs propres.

Soit A une matrice symétrique. Admettons que l'un de ses vecteurs propres p est déjà obtenu et normé, de sorte que $\| p \| = 1$. Si l'on veut construire la suite convergeant vers un autre vecteur propre, il est naturel de choisir pour vecteur initial un vecteur orthogonal à p. Mais le vecteur x_1 obtenu après la première itération de la méthode des puissances ne sera plus orthogonal à p et l'on devra le rectifier. Cela engendre le processus suivant. Si le vecteur x_k est construit, on pose

$$y_{k+1} = Ax_k, \quad z_{k+1} = y_{k+1} - p'py_{k+1}, \quad \alpha_{k+1}x_{k+1} = z_{k+1}.$$

Le facteur α_{k+1} est choisi de manière que soit remplie la condition $\|x_{k+1}\| = 1$. Il est facile de voir que x_{k+1} est orthogonal à p. En réunissant les égalités précédentes, on obtient

$$\alpha_{k+1}x_{k+1} = (E - p\,'p)Ax_k.$$

Ainsi donc, la construction de la suite équivaut à l'application de la méthode des puissances de matrice

$$B = (E - p\,'p)A.$$

Cela signifie que la suite $\{x_k\}$ converge vers le vecteur propre q de la matrice B associé au nombre caractéristique μ_1 maximal en module de cette matrice.

Pour trouver μ_1, notons que B a les mêmes nombres caractéristiques que A, mais la multiplicité de la racine λ_1 a diminué de 1, tandis que la multiplicité du nombre caractéristique nul a augmenté de 1 (ou est apparue une racine $\lambda = 0$ si elle n'existait pas). En effet, considérons une matrice orthogonale S dont la première colonne est p, et à l'aide de cette matrice formons la matrice $B' = S^{-1}BS$. On peut l'écrire ainsi $[S^{-1}(E - p\,'p)S][S^{-1}AS]$. Les facteurs renfermés entre crochets sont respectivement de la forme

$$\left\|\begin{matrix} 0 & \dots & 0 \\ \vdots & E_{n-1} & \\ 0 & & \end{matrix}\right\| \quad \text{et} \quad \left\|\begin{matrix} \lambda_1 & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \bar{A} & \\ 0 & & & \end{matrix}\right\|.$$

En multipliant ces matrices, on obtient

$$B' = \left\|\begin{matrix} 0 & \dots & 0 \\ \vdots & \bar{A} & \\ 0 & & \end{matrix}\right\|.$$

Les $n - 1$ dernières lignes dans la matrice B' sont les mêmes que dans le second facteur. En comparant les polynômes caractéristiques des matrices A, $\bar{A}$ et B, on aboutit au résultat exigé.

Ainsi, $\mu_1 = \lambda_1$ si la multiplicité de λ_1 dans le polynôme caractéristique de la matrice A est strictement supérieure à l'unité ; dans le cas contraire, μ_1 est égal au nombre caractéristique de la matrice A, de module immédiatement inférieur.

On voit aisément que le vecteur q est orthogonal à p. En effet,

$$\mu_1\,'pq = {}'pBq = {}'p(E - p\,'p)Aq = \mathbf{0}Aq = 0.$$

D'où pour $\mu_1 \neq 0$, on obtient $'pq = 0$.

Démontrons que pour une matrice symétrique A le vecteur q sera de même son vecteur propre. En effet, le produit d'un vecteur par la matrice $E - p'p$ est égal à la projection orthogonale de ce vecteur sur le supplémentaire orthogonal $\mathcal{N}$ du sous-espace engendré par p. Aussi l'égalité $(E - p'p)Aq = \mu_1 q$ signifie-t-elle que

$$Aq = \mu_1 q + \alpha p. \tag{6}$$

Il en découle que si $\mathcal{N}$ est invariant par A (en particulier, pour une matrice symétrique), on a $Aq = \mu_1 q$, ce qu'il fallait démontrer.

Ainsi, dans le cas de matrice symétrique, la suite de x_k converge vers le vecteur propre de A, othogonal à p.

L'application de la méthode des puissances à la matrice B peut être simplifiée si on réduit B à la matrice B' indiquée plus haut, dans laquelle la première ligne et la première colonne sont nulles. Toute puissance de la matrice B' est de la même forme et, par suite, tout se réduit à l'application de la méthode des puissances à la matrice $\tilde{A}$ d'ordre $n - 1$. On ne s'arrêtera pas à la réalisation de cette méthode d'abaissement de l'ordre ainsi que sur les autres méthodes semblables appelées *méthodes d'exhaustion*.

Supposons maintenant que la matrice A n'est pas symétrique. En opérant comme il a été indiqué plus haut, on peut construire le vecteur q, mais dans la relation (6) le nombre α peut s'avérer différent de zéro. Si la multiplicité de la racine λ_1 est strictement supérieure à 1, il faut pour cela que le vecteur p possède le premier vecteur associé. En effet, pour $\mu_1 = \lambda_1$ et $\alpha \neq 0$ la relation (6) peut être mise sous la forme

$$(A - \lambda_1 E)(\alpha^{-1} q) = p.$$

Il n'est pas difficile de construire un exemple dans lequel q est associé au vecteur p construit auparavant.

Si la racine λ_1 n'est pas multiple, $\mu_1 \neq \lambda_1$, on peut d'après q construire le vecteur propre r de la matrice A, associé au nombre caractéristique μ_1. On cherchera ce vecteur sous la forme $r = q + \beta p$. On vérifie aisément que l'égalité

$$A(q + \beta p) = \mu_1(q + \beta p)$$

est vérifiée pour $\beta = \alpha/(\mu_1 - \lambda_1)$.

Supposons que les nombres caractéristiques de la matrice A sont tels que $|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots$, avec λ_1 et λ_2 réels. Pour trouver λ_2 et le vecteur propre correspondant on peut profiter alors de la méthode suivante. Souvenons-nous que chaque vecteur propre de la transformation adjointe $\mathbf{A}^*$ est orthogonal à tous les vecteurs propres de $\mathbf{A}$ associés à d'autres valeurs propres. En appliquant la méthode des puissances directe, on peut obtenir un vecteur unitaire u pour lequel $'Au = \lambda_1 u$, $'up = \sigma \neq 0$

et ${}^t uq = 0$, où p et q sont les vecteurs propres de A associés respectivement à λ_1 et λ_2. On peut montrer que le processus construit suivant les formules

$$y_{k+1} = Ax_k, \quad z_{k+1} = y_{k+1} - u\,{}^t u y_{k+1}, \quad \alpha_{k+1} x_{k+1} = z_{k+1},$$

converge vers le vecteur propre associé à λ_2. La démonstration diffère peu de celle qui a été fournie plus haut pour la matrice symétrique et on laissera au soin du lecteur de la réaliser.

Revenons au cas de la matrice symétrique A et posons que nous avons construit deux de ses vecteurs propres $p^{(1)}$ et $p^{(2)}$. Le troisième vecteur propre de cette matrice peut être construit de façon analogue par orthogonalisation à chaque itération du vecteur Ax_k par rapport à $p^{(1)}$ et $p^{(2)}$. D'une façon générale, si on a construit le système orthonormé de s vecteurs propres $p^{(1)}, \ldots, p^{(s)}$, le vecteur suivant peut être construit d'après la méthode des puissances appliquée à la matrice

$$\left(E - \sum_{i=1}^{s} (p^{(i)})\,{}^t(p^{(i)})\right) A,$$

ce qui équivaut à l'orthogonalisation du vecteur Ax_k par rapport à tous les $p^{(1)}, \ldots, p^{(s)}$. On peut ainsi obtenir une base orthonormée de vecteurs propres de toute matrice symétrique.

Une autre méthode proche de la méthode décrite ne diffère que par le fait que les vecteurs propres de la matrice sont calculés non pas successivement mais en même temps. Dans ce cas, les vecteurs construits à chaque itération constituent les colonnes d'une même matrice, de sorte qu'on construit une suite de matrices.

Le processus correspondant à la méthode des puissances directe sans décalage est construit suivant les formules

$$AP_k = V_{k+1}, \quad V_{k+1} = P_{k+1} R_{k+1}, \tag{7}$$

où P_k est une matrice à colonnes orthonormées, et R_{k+1} est choisie de manière que soient orthonormées les colonnes de P_{k+1}.

Dans la suite de l'exposé, il nous faudra de plus en plus passer sous silence certains détails. En particulier, on n'étudiera pas la convergence du processus (7).

Par ce processus, on peut en particulier obtenir un couple de racines λ_1 et $\bar{\lambda}_1$ complexes conjugées, ainsi qu'un sous-espace bidimensionnel invariant correspondant, si ces racines dépassent en module les autres nombres caractéristiques. Dans ce cas, les matrices P_k sont constituées de deux colonnes qui tendent vers les vecteurs de la base orthonormée du sous-espace invariant. R_k sont les matrices d'ordre deux dont les nombres caractéristiques tendent vers les nombres caractéristiques complexes conjugués λ_1 et $\bar{\lambda}_1$ de la matrice A.

Si on ne connaît rien sur les nombres caractéristiques de la matrice A, il vaut mieux réaliser le processus (7) en utilisant les matrices P_k à n colonnes orthonormées, c'est-à-dire des matrices orthogonales. Dans ce cas, P_{k+1} et R_{k+1} *peuvent être obtenues par la* QR-décomposition de la matrice A^k.

En réunissant les formules (7), on peut, comme dans les cas précédents, écrire le résultat de la k-ième itération :

$$A^k P_0 = P_k R_k R_{k-1} \dots R_1.$$

Si l'on débute le processus par une matrice unité et qu'on introduise la notation $U_k = R_k R_{k-1} \dots R_1$, il vient

$$A^k = P_k U_k.$$

La matrice U_k étant une matrice triangulaire supérieure, la dernière formule peut être prise pour une QR-décomposition de A^k.

Ce processus permet d'obtenir tous les nombres caractéristiques et, au moins pour une matrice symétrique, tous les vecteurs propres, mais en général, il revêt une autre forme, beaucoup plus commode. C'est à sa description qu'on passe.

5. *QR*-algorithme. Cet algorithme s'appuie sur le processus suivant. Recherchons la QR-décomposition de la matrice initiale. Soit $A = Q_1 R_1$. Posons $A_1 = R_1 Q_1$ et cherchons pour la matrice A_1 sa QR-décomposition, soit $A_1 = Q_2 R_2$. On obtient la matrice A_2 en permutant les facteurs Q_2 et R_2. D'une façon générale,

$$A_{k-1} = Q_k R_k, \quad A_k = R_k Q_k. \tag{8}$$

Ceci étant, $R_k = Q_k^{-1} A_{k-1}$ et par suite, $A_k = Q_k^{-1} A_{k-1} Q_k$. D'où

$$A_k = Q_k^{-1} \dots Q_1^{-1} A Q_1 \dots Q_k = P_k^{-1} A P_k, \tag{9}$$

où $P_k = Q_1 \dots Q_k$. Cela montre que les nombres caractéristiques de toutes les matrices A_k coïncident avec les nombres caractéristiques de la matrice A.

Si l'on désigne le produit $R_k \dots R_1$ par U_k, la puissance k-ième de la matrice A peut être exprimée ainsi :

$$A^k = P_k U_k. \tag{10}$$

En effet,

$$P_k U_k = Q_1 \dots Q_k R_k \dots R_1 = Q_1 \dots Q_{k-1} A_{k-1} R_{k-1} \dots R_1 = P_{k-1} A_{k-1} U_{k-1}.$$

Or la formule (9) écrite pour A_{k-1} implique

$$P_{k-1} A_{k-1} = A P_{k-1},$$

et, par suite, $P_k U_k = A P_{k-1} U_{k-1}$. On obtient la formule (10) par application successive de ce résultat, ce qui permet d'établir le lien avec le processus décrit à la fin du point précédent.

On voudrait affirmer que la suite de matrices A_k converge vers une matrice triangulaire à blocs, mais ce n'est malheureusement pas le cas. On peut seulement démontrer que les éléments de A_k situés au-dessous des blocs diagonaux tendent vers zéro, tandis que les éléments de ces blocs et les éléments situés ci-dessus sont uniformément bornés. La suite de matrices ayant cette propriété est dite *convergente en forme* vers la matrice triangulaire à blocs.

A ce qu'il paraît, il n'existe pas jusqu'à présent de démonstration de cette assertion aussi simple pour qu'on puisse la reproduire ici.

Pour décrire plus en détail la matrice « limite » A_∞ obtenue à l'aide du QR-algorithme, supposons que les nombres caractéristiques de A sont ordonnés de façon que leurs modules forment une suite décroissante :

$$|\lambda_1| = \ldots = |\lambda_{r_1}| > |\lambda_{r_1+1}| = \ldots$$
$$\ldots = |\lambda_{r_1+r_2}| > \ldots > |\lambda_{n-r_m}| = \ldots = |\lambda_n|.$$

La matrice A_k peut être partagée en blocs correspondant à des groupes de nombres caractéristiques égaux en module. On arrive à démontrer que les éléments du (i, j)-ième bloc au-dessous de la diagonale tendent pour $i > j$ vers zéro non moins vite que

$$O\left(k^{2(m-1)}\left|\frac{\lambda_{r_j}}{\lambda_{r_i}}\right|^k\right), \tag{11}$$

où m est la multiplicité maximale des racines dans le polynôme minimal de la matrice A (voir Voïévodine [40]).

Dans la matrice « limite », les blocs diagonaux correspondent aux groupes de nombres caractéristiques égaux en module. Ceci étant, les nombres caractéristiques des blocs se confondent avec les nombres caractéristiques de la matrice A associés au groupe correspondant. Par exemple, à un couple de nombres complexes conjugués λ, $\bar{\lambda}$ de multiplicité 1 correspond un bloc diagonal d'ordre deux avec nombres caractéristiques λ et $\bar{\lambda}$.

Admettons que la matrice A possède un système complet de vecteurs propres et que X est une matrice dont les colonnes sont les vecteurs propres de la matrice A. Notons Q_∞ la matrice orthogonale qui intervient dans la QR-décomposition de la matrice X. On peut alors démontrer qu'il existe des matrices diagonales orthogonales T_k telles que $T_k Q_k \to Q_\infty$.

Cette relation est particulièrement importante pour une matrice symétrique A car elle permet de trouver sa base orthonormée de vecteurs propres.

Habituellement, les dimensions des blocs diagonaux de la matrice limite A_∞ ne sont pas grandes, et les nombres caractéristiques correspondants s'obtiennent sans difficulté. Une fois ceci fait, les vecteurs propres associés sont recherchés par la méthode des puissances inverse avec décalage.

6. Réduction de la matrice à la forme quasi triangulaire. Chaque étape du QR-algorithme comporte la recherche de la QR-décomposition d'une matrice et la multiplication de matrices, de sorte qu'elle exige un nombre considérable d'opérations arithmétiques. On peut faciliter sensiblement les calculs si on prépare au préalable la matrice A en la réduisant à la forme quasi triangulaire (ou forme de Hessenberg).

La matrice A d'éléments a_{ik} est par définition une matrice *quasi triangulaire* supérieure si $a_{ij} = 0$ pour $i > j + 1$. Ainsi, les seuls éléments non nuls situés au-dessous de la diagonale principale sont ceux da la rangée parallèle à la diagonale et située immédiatement au-dessous d'elle. De façon analogue on détermine les matrices quasi triangulaires inférieures.

PROPOSITION 1. *Pour toute matrice A on peut construire une matrice orthogonale S telle que la matrice $S^{-1}AS$ soit une matrice quasi triangulaire supérieure.*

Pour réduire la matrice donnée à une matrice quasi triangulaire supérieure, on procède à chaque étape à la transformation d'une colonne de la matrice à la forme exigée. Supposons qu'après $r - 1$ étapes la matrice A s'est transformée en une matrice $A^{(r)}$ dont les $r - 1$ premières colonnes correspondent à la forme quasi triangulaire supérieure. Alors la matrice $A^{(r+1)}$ se déduit d'après la formule

$$A^{(r+1)} = {}^tP_rA^{(r)}P_r,$$

où P_r est une matrice de symétrie (point 6, § 3). La matrice de symétrie est symétrique et orthogonale : $P_r = {}^tP_r = P_r^{-1}$, de sorte que $A^{(r+1)} = P_r^{-1}A^{(r)}P_r$.

Décrivons comment on choisit la matrice P_r. Si dans la r-ième colonne de la matrice $A^{(r)}$ les $n - r$ derniers éléments sont des zéros, on n'a pas besoin de transformer cette colonne et on passe à la $(r + 1)$-ième colonne. Dans le cas contraire, on construit une symétrie qui transforme le vecteur

$$u = {}^t\|0, \ldots, 0, \alpha_{r+1,r}^{(r)}, \ldots, \alpha_{nr}^{(r)}\|$$

en un vecteur proportionnel à la $(r + 1)$-ième colonne de la matrice unité. Remarquons que ce choix diffère de celui qu'on a fait dans la construction de la QR-décomposition par la méthode des symétries, à savoir : on a substitué des zéros dans la r-ième colonne à $r - 1$ premiers éléments et non pas à r, après quoi on a transformé la colonne obtenue en une colonne proportionnelle à la r-ième colonne de la matrice unité. En définitive, on arrive à construire la matrice de symétrie P_r dont les r premières colonnes et lignes se confondent avec les lignes et colonnes respectives de la matrice unité.

La multiplication à gauche de $A^{(r)}$ par P_r laisse inchangées les r premières lignes de la matrice $A^{(r)}$, ainsi que les éléments nuls des colonnes de numéros $< r$, mais elle annulle les $n - (r + 1)$ derniers éléments de la

r-ième colonne. La multiplication à droite de la matrice par P_r ne modifie pas ses r premières colonnes et, par suite, la forme du produit à gauche se conserve.

La dernière colonne à transformer est la $(n - 2)$-ième. Donc, après $n - 2$ étapes la matrice A sera convertie en matrice quasi triangulaire supérieure.

PROPOSITION 2. *La forme de la matrice quasi triangulaire supérieure reste inchangée par toutes les transformations de la matrice suivant le QR-algorithme.*

DÉMONSTRATION. Soit A une matrice quasi triangulaire supérieure. Elle peut être transformée en une matrice triangulaire supérieure R si on la multiplie à gauche par les matrices de rotations dans les plans engendrés par les couples de vecteurs de base de numéros (1, 2), (2, 3), (3, 4), ..., $(n - 1, n)$. Donc, la QR-décomposition de A est de la forme

$$A = {}^tP_{12} \dots {}^tP_{n-1,\,n}R,$$

et la matrice A_1 est égale à

$$A_1 = RQ = R\,{}^tP_{12} \dots {}^tP_{n-1,\,n}.$$

La multiplication à droite par les matrices de rotation mentionnées ne peut modifier dans la matrice R que les blocs d'ordre deux qui sont situés sur la diagonale et au-dessus d'elle. Dans ce cas, la matrice triangulaire ne peut être convertie qu'en une matrice quasi triangulaire.

PROPOSITION 3. *La nature symétrique de la matrice A n'est pas perturbée par l'application à cette dernière du QR-algorithme.*

En effet, si A est symétrique et $A = QR$, on a $A = {}^tA = {}^tR\,{}^tQ$, d'où ${}^tR = QRQ$. Maintenant si $A_1 = RQ$, on obtient ${}^tA_1 = {}^tQ\,{}^tR = {}^tQQRQ = A_1$, ce qu'il fallait démontrer.

Si la matrice symétrique a été transformée par une matrice orthogonale S en une forme quasi triangulaire, la matrice obtenue $S^{-1}AS$ est aussi symétrique. Les matrices quasi triangulaires symétriques sont dites *tridiagonales*. Dans une telle matrice, les seuls éléments non nuls sont ceux de la diagonale principale et de deux rangées qui lui sont parallèles et situées tout près d'elle, l'une au-dessus et l'autre au-dessous.

Le QR-algorithme est pratiquement toujours appliqué aux matrices quasi triangulaires et, dans le cas de matrices symétriques, aux matrices tridiagonales.

Remarquons que la QR-décomposition d'une matrice quasi triangulaire est obtenue le plus aisément par la méthode des rotations : comme on l'a vu en démontrant la proposition 2, il ne faut que $n - 1$ rotations.

7. Accélération de la convergence du QR-algorithme. L'estimation (11)

donnée plus haut montre que la vitesse de convergence du QR-algorithme décrit peut s'avérer très faible. On a mis au point des variantes améliorées de cet algorithme pour permettre d'augmenter sensiblement la vitesse de convergence.

a) *Abaissement de l'ordre*. Si au cours de transformations d'une matrice quasi triangulaire d'après le QR-algorithme un de ses éléments situés au-dessous de la diagonale s'avère si petit qu'on peut le négliger et remplacer par le zéro, la matrice devient une matrice triangulaire à blocs diagonaux quasi triangulaires. On obtient tous les nombres caractéristiques de la matrice initiale si l'on trouve les nombres caractéristiques des blocs diagonaux. Aussi l'apparition des éléments nuls au-dessous de la diagonale diminue-t-elle les dimensions du problème à résoudre.

b) *Décalages*. Soit une suite numérique $\{\rho_k\}$ appelée *suite de décalages*. Le QR-algorithme avec décalage est défini par les formules

$$A_{k-1} - \rho_k E = Q_k R_k, \quad A_k = R_k Q_k + \rho_k E.$$

Il n'est pas difficile de démontrer que dans ce cas $A_k = {}^tP_k A P_k$, où $P_k = Q_1 \ldots Q_k$. Ainsi, les nombres caractéristiques de toutes les matrices A_k coïncident avec les nombres caractéristiques de la matrice A.

On peut également montrer que

$$P_k U_k = (A - \rho_1 E) \ldots (A - \rho_k E).$$

On a mis au point une série de critères qui permettent de choisir des suites de décalages. On ne les exposera pas. Notons seulement que la plus grande accélération est fournie par un décalage proche du nombre caractéristique de la matrice A. On rencontre des cas où la vitesse de convergence est anormalement faible. Il existe également des matrices qui restent invariantes par toutes les transformations du QR-algorithme (l'unique bloc diagonal de la matrice limite est confondu avec la matrice de départ). Dans ces situations, tout décalage peut s'avérer utile.

Il existe aussi de nombreuses complications et modifications du QR-algorithme avec décalage, adaptées aux différentes situations particulières, par exemple, à la recherche des nombres caractéristiques conjugués complexes d'une matrice réelle.

8. Estimations *a posteriori* de la précision des calculs. Si on a calculé une valeur approchée de la solution du problème complet des valeurs propres, il est possible d'utiliser le résultat obtenu pour estimer sa précision dans une large classe de problèmes. On admettra que les nombres caractéristiques calculés $\lambda_1^*, \ldots, \lambda_n^*$ sont différents et que les vecteurs propres calculés $\boldsymbol{\xi}_1^*, \ldots, \boldsymbol{\xi}_n^*$ sont linéairement indépendants.

On peut former n vecteurs-écarts

$$\boldsymbol{\rho}_i = A\boldsymbol{\xi}_i^* - \lambda_i^* \boldsymbol{\xi}_i^*.$$

En désignant par X et P les matrices composées respectivement à partir des colonnes ξ_i^* et ρ_i, mettons ce système d'égalités sous la forme

$$AX - XD = P,$$

avec $D = \operatorname{diag}(\lambda_1^*, \ldots, \lambda_n^*)$. En multipliant par X^{-1}, on obtient

$$X^{-1}AX = D + X^{-1}P.$$

Cette égalité signifie que la transformée exacte de la matrice A par X est une matrice qui diffère de la matrice diagonale par $V = X^{-1}P$. Si λ_i^* et ξ_i^* sont de bonnes approximations, la norme de la matrice P est petite. Si $\| V \|$ est aussi petite, les nombres caractéristiques de $D + V$ peuvent être estimés d'après les nombres caractéristiques de D, c'est-à-dire que les nombres caractéristiques exacts de A sont estimés d'après leurs approximations calculées. Cette estimation peut être réalisée par construction des disques de localisation (voir § 4, ch. XII) si les éléments de la matrice V sont calculés de façon assez précise. Pour aboutir à la précision maximale, la matrice d'écarts P est calculée en régime d'accumulation du produit scalaire, et les solutions du système d'équations linéaires, au moyen duquel on obtient les colonnes de la matrice V, sont soumises à une précision itérative (point 3, § 4).

Supposons maintenant qu'on a calculé un seul nombre caractéristique λ_1^* et le vecteur propre associé ξ_1^*. Voyons comment on peut estimer leur précision. Il est naturel de recommencer par l'écart

$$A\xi_1^* - \lambda_1^*\xi_1^* = \rho.$$

Sans restreindre la généralité on peut considérer que $\|\xi_1^*\|^2 = {}'\xi_1^*\xi_1^* = 1$, et mettre l'égalité précédente sous la forme

$$(A - \lambda_1^*E - \rho'\xi_1^*)\xi_1^* = 0.$$

Cela montre que λ_1^* est la valeur propre et que ξ_1^* est le vecteur propre de la matrice perturbée $A + H$, où $H = -\rho'\xi_1^*$. *Ceci étant, il n'est pas difficile de montrer que* $\| H \| = \|\rho\|$ (voir p. 441). Selon la formule (14) du § 2, il vient

$$\min |\lambda_i - \lambda_1^*| \leqslant \nu(A)\|\rho\|.$$

Cette estimation est peu utile. D'abord, $\nu(A)$ est inconnu. Il est possible de le majorer si l'on trouve les autres vecteurs propres et le nombre conditionnel de la matrice qu'ils composent. En outre, cette estimation peut s'avérer exagérée si la première valeur propre de la matrice est bien conditionnée et qu'il existe des valeurs mal conditionnées.

Considérons une autre approche. Soient ξ_1^* et η_1^* les approximations calculées des vecteurs propres des matrices A et ${}'A$ respectivement et soient les vecteurs exacts ξ_1 et η_1 associés à une même valeur propre λ_1. On

admettra que les vecteurs $\boldsymbol{\xi}_1$ et $\boldsymbol{\eta}_1$ sont normés et que les différences $\delta\boldsymbol{\xi}_1 = \boldsymbol{\xi}_1^* - \boldsymbol{\xi}_1$ et $\delta\boldsymbol{\eta}_1 = \boldsymbol{\eta}_1^* - \boldsymbol{\eta}_1$ appartiennent aux enveloppes linéaires des vecteurs $\boldsymbol{\xi}_2, \ldots, \boldsymbol{\xi}_n$ et $\boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_n$ respectivement (voir p. 394). En vertu de cette convertion $'\boldsymbol{\eta}_1\delta\boldsymbol{\xi}_1 = 0$ et $\delta'\boldsymbol{\eta}_1\boldsymbol{\xi}_1 = 0$. Calculons l'expression

$$\frac{'\boldsymbol{\eta}_1^* A \boldsymbol{\xi}_1^*}{'\boldsymbol{\eta}_1^* \boldsymbol{\xi}_1^*} = \frac{('\boldsymbol{\eta}_1 + \delta'\boldsymbol{\eta}_1^*)(\lambda_1\boldsymbol{\xi}_1 + A\delta\boldsymbol{\xi}_1)}{s_1^*} = \lambda_1 + \frac{\delta\boldsymbol{\eta}_1(A - \lambda_1 E)\delta\boldsymbol{\xi}_1}{s_1^*}, \quad (12)$$

où $s_1^* = {'\boldsymbol{\eta}_1^*}\boldsymbol{\xi}_1^* = s_1 + \delta'\boldsymbol{\eta}_1\delta\boldsymbol{\xi}_1$. On voit que le coefficient calculé s_1^* diffère de la valeur exacte s_1 par une grandeur dont l'ordre est le produit d'erreurs. Si ce coefficient n'est pas trop petit, l'expression calculée diffère très peu de la vraie valeur propre.

Si la matrice A est symétrique, $\boldsymbol{\xi}_1$et $\boldsymbol{\eta}_1$ peuvent être pris confondus et l'expression (12) devient le quotient de Rayleigh pour la matrice A. En raison de ce fait, l'expression considérée est appelée *quotient de Rayleigh généralisé*. Pour les matrices symétriques, le quotient permet de simplifier les calculs et d'aboutir à des bonnes estimations théoriques. Dans le cas des matrices non symétriques on peut seulement dire que, si $\mu = \frac{'\boldsymbol{\eta}_1^* A \boldsymbol{\xi}_1^*}{s_1^*}$ fournit, avec $\boldsymbol{\eta}_1^*$ et $\boldsymbol{\xi}_1^*$, des écarts faibles, il est une bonne approximation de λ_1, et s_1^* est une bonne approximation de s_1.

La valeur μ du quotient de Rayleigh généralisé peut être utilisée non seulement pour estimer l'erreur mais aussi pour préciser le résultat. A cet effet, en prenant μ pour la valeur de décalage, on obtient le vecteur propre par la méthode des puissances inverse.

CHAPITRE XIV

PSEUDO-SOLUTIONS ET MATRICES PSEUDO-INVERSES

§ 1. Propriétés élémentaires

1. Remarques préliminaires. Dans les problèmes pratiques il est souvent nécessaire de trouver une solution vérifiant un grand nombre de contraintes parfois contradictoires. Si un tel problème se réduit à un système d'équations linéaires, ce système s'avère généralement incompatible. Dans ce cas, le problème ne peut être résolu que par le choix d'un accommodement : on essaie de vérifier presque toutes les contraintes. Expliquons-le sur l'exemple suivant.

Supposons que deux grandeurs physiques y et x sont liées dans le domaine de leur variation par une dépendance linéaire de la forme $y = kx + b$ et que les coefficients doivent être établis expérimentalement. Les données expérimentales sont constituées par m points du plan de coordonnées : (x_1, y_1), ..., (x_m, y_m).

Si ces couples de valeurs sont en effet liés par la dépendance recherchée, en les portant dans l'équation on aboutit à un système de m équations linéaires à deux inconnues k et b :

$$y_i = kx_i + b, \quad i = 1, \dots, m.$$

Pour tous x_i et x_j différents, les points (x_i, y_i) et (x_j, y_j) définissent une seule droite. Mais chaque couple de points définit sa propre droite, et on n'a aucune raison de préférer une droite à une autre.

Si les données expérimentales méritent une certaine confiance, l'incompatibilité du système sert de raison pour rejeter l'hypothèse de la dépendance linéaire. Le problème de compatibilité des données expérimentales avec l'hypothèse de dépendance linéaire est résolu par l'analyse statistique.

Supposons que la précision de l'information initiale admet l'existence de la dépendance linéaire. Dans ce cas, la chose qu'on doit faire en réalité, c'est obtenir une droite du plan de coordonnées qui soit le plus près de tous les points expérimentaux, sans passer peut-être par aucun couple de ces points, ni même par aucun de ces points (fig. 55).

Pour pouvoir dire si un point est proche de la droite on utilise en général dans ce problème non pas la distance du point à la droite mais la diffé-

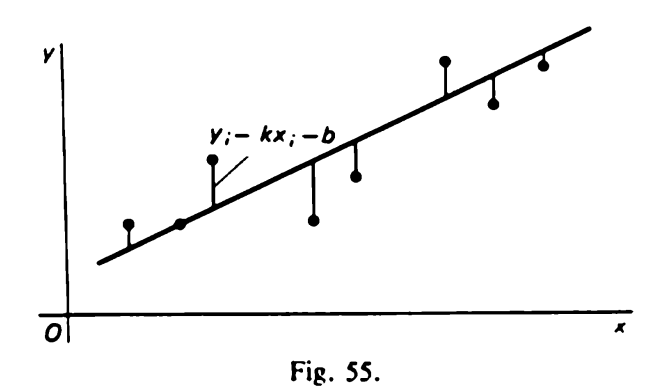


Fig. 55.

rence des ordonnées $y_i - kx_i - b$ et l'on choisit une droite de manière que la somme des carrés de toutes ces différences soit minimale. Les coefficients k_0 et b_0 de l'équation de cette droite définissent la solution du problème posé, qui n'est en aucune sorte la solution du système d'équations linéaires (ne présentant en général aucune solution). On peut prendre k_0 et b_0 pour solution généralisée du système ou, comme on dit, sa *pseudo-solution*. La définition exacte de cette notion sera donnée plus loin.

On se limitera au § 1 aux propriétés élémentaires des pseudo-solutions et des matrices pseudo-inverses associées, qu'on peut déduire facilement sans recourir à autre chose qu'aux théorèmes sur les systèmes d'équations linéaires et les supplémentaires orthogonaux des sous-espaces euclidiens, connus du cours général. Ce paragraphe peut ainsi être assimilé par le lecteur indépendamment des chapitres précédents.

On envisagera le système d'équations linéaires

$$Ax = b \tag{1}$$

à matrice A de type (m, n). La lettre r désignera le rang de cette matrice. Aucune contrainte n'est imposée en général à m, n et r. Puisque x est une matrice-colonne à n éléments, et b une matrice-colonne à m éléments, il est naturel de se servir, pour l'interprétation géométrique, des espaces arithmétiques $\mathscr{R}_n$ et $\mathscr{R}_m$. Par norme de la matrice-colonne x d'éléments x^1, ..., x^n on entendra sa norme euclidienne, c'est-à-dire le nombre

$$\|x\| = \sqrt{{}^t xx} = \sqrt{(x^1)^2 + \ldots + (x^n)^2}.$$

2. Minimisation de l'écart. On appelle *écart* engendré par la matrice-colonne x après sa substitution dans le système d'équations (1) la matrice-colonne

$$u = b - Ax.$$

La solution de système est la matrice-colonne qui donne un écart nul.

Si le système (1) est incompatible, il est naturel de rechercher une matrice-colonne x dont l'écart a une norme minimale. Si une telle matrice-colonne existe, elle peut être considérée comme solution généralisée du

système. Il va de soi que si le système est compatible, sa solution est également une solution généralisée.

En rapport avec ce qui vient d'être dit notons le fait suivant. On a déjà étudié les écarts au ch. XIII et on a essayé à diminuer leur normes, mais la situation y était différente. Le système avait une solution et la matrice-colonne à faible écart était prise pour approximation de cette solution. Tandis qu'ici on ne sait pas si la solution existe et on s'intéresse â la matrice-colonne présentant un écart minimal. On étudie maintenant de façon théorique les solutions généralisées en supposant que les calculs peuvent être effectués exactement. Or dans la pratique, la solution généralisée, comme d'ailleurs tout autre objet, ne peut être calculée que de façon approchée.

Pour comparer les écarts, profitons de la norme euclidienne et recherchons donc la matrice-colonne x pour laquelle est minimale la grandeur

$$\|u\|^2 = {}'(b - Ax)(b - Ax). \tag{2}$$

En considérant les éléments de la matrice-colonne x comme des variables indépendantes, calculons la différentielle totale de $\|u\|^2$. Il n'est pas difficile de vérifier que

$$\mathrm{d}\|u\|^2 = -\mathrm{d}'x\,'A(b - Ax) - {}'(b - Ax)A\,\mathrm{d}x.$$

Vu que le second terme est une matrice d'ordre un qui ne varie pas par transposition, il vient

$$\mathrm{d}\|u\|^2 = -2\mathrm{d}\,'x\,'A(b - Ax).$$

Aussi la différentielle s'annule-t-elle si et seulement si

$$'AAx = {}'Ab. \tag{3}$$

Ce système d'équations linéaires associé au système (1) est dit *normal.* Il résulte du système (1). Indépendamment de la compatibilité du système (1), on peut énoncer la

PROPOSITION 1. *Le système d'équations normal est obligatoirement compatible.*

Cette proposition est un nouvel énoncé de la proposition 12 du § 1, ch. XI. Toutefois, la démonstration directe en termes de matrices n'est pas compliquée et on va la fournir. La matrice $'AA$ est symétrique, de sorte que le système homogène transposé du système (3) est de la forme $'AAy = \mathbf{0}$. Pour toute solution de ce système, on a les égalités découlant successivement l'une de l'autre :

$$'y\,'AAy = {}'(Ay)(Ay) = 0, \quad Ay = \mathbf{0}, \quad 'y('Ab) = 0.$$

La dernière d'entre elles signifie que le système (3) vérifie l'hypothèse du théorème de Fredholm, ce qui achève la démonstration.

PROPOSITION 2. *La borne inférieure du carré de la norme de l'écart est atteinte pour les seules solutions du système normal* (3).

DÉMONSTRATION. Ecrivons la formule (2) pour la matrice-colonne $x_0 + \Delta x$ et ouvrons les parenthèses. Il vient

$$'(b - A(x_0 + \Delta x))(b - A(x_0 + \Delta x)) =$$
$$= '(b - Ax_0)(b - Ax_0) - 2\,'(\Delta x)\,'A(b - Ax_0) + '(\Delta x)\,'AA\Delta x.$$

Le dernier terme est positif car $'(\Delta x)'AA\Delta x = '(A\Delta x)(A\Delta x) \geqslant 0$. Si x_0 vérifie le système (3), le second terme est nul et alors l'addition de Δx ne diminue pas la valeur de la fonction, quelle que soit cette matrice-colonne Δx.

Inversement, la fonction définie pour tous les x n'atteint sa borne inférieure qu'en un point d'extrémum local. Or en ces points la différentielle est nulle, si bien que la condition (3) est satisfaite. La proposition est démontrée.

Selon le théorème connu (théorème 2, § 5, ch. V), l'ensemble de toutes les solutions du système normal peut être décrit par la formule $x = x_0 + z$, où x_0 est une solution fixée du système normal et z une solution quelconque du système homogène $'AAz = \mathbf{0}$. On a vu en démontrant la proposition 1 que ce dernier système est équivalent à $Az = \mathbf{0}$, de sorte qu'on peut poser que la solution du système normal (3) est définie à une solution quelconque près du système homogène $Az = \mathbf{0}$. On aboutit ainsi à la

PROPOSITION 3. *Le système normal possède une solution unique si et seulement si le système $Az = \mathbf{0}$ n'a qu'une solution triviale, c'est-à-dire si les colonnes de la matrice A sont linéairement indépendantes. En particulier, ce sera rempli si la matrice A est régulière.*

Si la solution du système normal n'est pas unique, il devient nécessaire de choisir l'une des solutions et l'on choisit la solution dont la norme est minimale.

DÉFINITION. On appelle *pseudo-solution normale* d'un système d'équations linéaires la matrice-colonne de norme minimale parmi toutes les matrices-colonnes qui, une fois portées dans ce système, donnent un écart minimal en norme.

Puisqu'il n'y a pas de risque d'ambiguïté, on appellera parfois la pseudo-solution normale tout simplement pseudo-solution.

Démontrons l'existence et l'unicité de la pseudo-solution normale. Pour le faire, il est commode d'utiliser les résultats de la théorie des espaces euclidiens. Considérons les espaces de matrices-colonnes $\mathscr{R}_n$ et $\mathscr{R}_m$ munis du produit scalaire défini par la formule $(x, y) = {}'xy$.

Soit $\mathscr{K} \subseteq \mathscr{R}_n$ l'ensemble des solutions du système homogène $Az = \mathbf{0}$ et soit $\mathscr{I} \subseteq \mathscr{R}_n$ l'ensemble des matrices-colonnes de la forme $'Ab$ pour tous

les $b \in \mathscr{R}_m$ possibles. La condition $p \in \mathscr{I}$ est équivalente à la compatibilité du système d'équations linéaires ${}^tAx = p$. D'autre part, selon le théorème de Fredholm, le dernier système est compatible si et seulement si pour chaque $z \in \mathscr{K}$ est vérifiée l'égalité ${}^tzp = 0$. Cela signifie que

$$\mathscr{K} = \mathscr{I}^{\perp}.$$

Rappelons que, quel que soit le sous-espace $\mathscr{L}$ dans l'espace euclidien, tout vecteur x peut être décomposé de façon unique en somme de la forme $x' + x''$, où $x' \in \mathscr{L}$ et $x'' \in \mathscr{L}^{\perp}$. Les vecteurs x' et x'' sont appelés *projections orthogonales* de x sur $\mathscr{L}$ et $\mathscr{L}^{\perp}$ respectivement.

Notons $\mathscr{N}$ l'ensemble des solutions du système normal d'équations linéaires (3). On a vu que les matrices-colonnes de $\mathscr{N}$ sont définies par la formule $x = x_0 + z$, où x_0 est une matrice-colonne de $\mathscr{N}$ et $z \in \mathscr{K}$.

THÉORÈME 1. *Tout système d'équations linéaires possède une pseudo-solution normale et une seule.*

En vertu de la proposition 2, il suffit de démontrer que dans l'ensemble $\mathscr{N}$ il existe toujours une matrice-colonne de norme minimale et une seule. Pour le démontrer, décomposons une matrice-colonne quelconque x de $\mathscr{N}$ en somme de ses projections orthogonales x_0 et x_1 sur $\mathscr{I}$ et sur $\mathscr{K}$.

Il est évident que x_0 appartient aussi à $\mathscr{N}$ parce que diffère de x par la matrice-colonne $x_1 \in \mathscr{K}$.

Si y est une autre matrice-colonne de $\mathscr{N}$, sa projection orthogonale sur $\mathscr{I}$ vaut également x_0. En effet, $y = (y - x + x_1) + x_0$, avec $(y - x) + x_1 \in \mathscr{K}$, et il reste à se référer à l'unicité des projections orthogonales. Ainsi donc, pour tous les vecteurs de $\mathscr{N}$ le terme x_0 est le même.

Pour un x quelconque de $\mathscr{N}$ on peut écrire

$$\|x\|^2 = {}^t(x_1 + x_0)(x_1 + x_0) = \|x_1\|^2 + \|x_0\|^2,$$

vu que ${}^tx_1x_0 = 0$. Il en découle que $\|x\| \geqslant \|x_0\|$, l'égalité étant réalisée si $x_1 = \mathbf{0}$, c'est-à-dire si $x = x_0$. Cela montre qu'il y a dans $\mathscr{N}$ une seule matrice-colonne dont la norme est minimale. Le théorème est démontré.

Il découle de la démonstration du théorème que la pseudo-solution normale peut être caractérisée par l'une quelconque des propriétés suivantes :

a) Elle est l'unique vecteur commun que possèdent $\mathscr{I}$ et $\mathscr{N}$:

$$x_0 = \mathscr{I} \cap \mathscr{N}, \tag{4}$$

c'est-à-dire qu'elle est l'unique solution du système normal, de forme

$$x_0 = {}^tAz. \tag{5}$$

b) Elle est la projection orthogonale de toute solution du système normal sur l'ensemble $\mathscr{I}$ des matrices-colonnes de la forme ${}^tAz$.

PROPOSITION 4. *Soient x_b et x_c les pseudo-solutions normales de deux*

systèmes d'équations linéaires $Ax = b$ et $Ax = c$. Alors $\beta x_b + \gamma x_c$ est la pseudo-solution normale du système $Ax = \beta b + \gamma c$.

DÉMONSTRATION. Il découle de ${}^tAAx_b = {}^tAb$ et de ${}^tAAx_c = {}^tAc$ que $\beta x_b + \gamma x_c$ vérifie le système normal : ${}^tAA(\beta x_b + \gamma x_c) = {}^tA(\beta b + \gamma c)$. Ensuite, il existe des matrices-colonnes z_b et z_c telles que $x_b = {}^tAz_b$ et $x_c = {}^tAz_c$. Donc, $\beta x_b + \gamma x_c = {}^tA(\beta z_b + \gamma z_c)$, ce qui achève la démonstration.

Il va de soi que la proposition 4 peut être étendue à toutes les combinaisons linéaires de matrices-colonnes.

Etudions quelques exemples fort simples.

1) Système de deux équations à une inconnue :

$$x = 1, \quad x = 2.$$

Le système normal associé à ce système est

$$\|1, 1\| \cdot \begin{Vmatrix} 1 \\ 1 \end{Vmatrix} x = \|1, 1\| \cdot \begin{Vmatrix} 1 \\ 2 \end{Vmatrix},$$

ou $2x = 3$. Il s'ensuit que la pseudo-solution vaut 3/2.

2) Système d'une équation à deux inconnues :

$$x^1 + x^2 = 2.$$

Le système d'équations normal est le système

$$\begin{Vmatrix} 1 \\ 1 \end{Vmatrix} \cdot \|1, 1\| \cdot \begin{Vmatrix} x^1 \\ x^2 \end{Vmatrix} = \begin{Vmatrix} 1 \\ 1 \end{Vmatrix} 2,$$

comprenant la même équation répétée deux fois. Sa solution générale est

$$\begin{Vmatrix} x^1 \\ x^2 \end{Vmatrix} = \begin{Vmatrix} 2 \\ 0 \end{Vmatrix} + \alpha \begin{Vmatrix} -1 \\ 1 \end{Vmatrix}.$$

La pseudo-solution sera la solution obtenue par multiplication de ${}^tA = \begin{Vmatrix} 1 \\ 1 \end{Vmatrix}$ par une matrice-colonne z à $m = 1$ éléments. On voit aisément que cette solution est

$$\begin{Vmatrix} x^1 \\ x^2 \end{Vmatrix} = \begin{Vmatrix} 1 \\ 1 \end{Vmatrix}.$$

3) Système d'une équation à une inconnue $\alpha x = \beta$. Si $\alpha \neq 0$, la pseudo-solution coïncide avec la solution $x = \beta/\alpha$. Mais si $\alpha = 0$, tout nombre x donne par substitution le même écart β de norme $|\beta|$. De tous les nombres il faut choisir celui dont la norme est minimale, c'est-à-dire 0. Ainsi donc, la pseudo-solution de l'équation $0x = \beta$ est 0.

On aurait obtenu le même résultat, s'il on avait complété la définition de la fonction $f(\alpha) = 1/\alpha$ par $f(0) = 0$. On voit que dans ce cas la pseudo-solution n'est pas une fonction continue par rapport aux éléments de la matrice du système et de la matrice-colonne des termes constants. Le même défaut est observé dans le cas général du système de m équations et n inconnues.

4) Système d'équations linéaires à matrice nulle $Ox = b$. La pseudo-solution est obtenue de la même façon que dans l'exemple 3 : tous les vecteurs engendrent le même écart égal à b, dont le vecteur nul qui a la norme minimale et qui est de ce fait la pseudo-solution du système.

3. Matrice pseudo-inverse. Pour une matrice carrée régulière A d'ordre n on peut définir la matrice inverse comme une matrice dont les colonnes sont solutions du système d'équations linéaires de la forme

$$Ax = e_i \tag{6}$$

où e_i est la i-ième colonne de la matrice unité d'ordre n. Par anologie, on peut donner la

DÉFINITION. On appelle *matrice pseudo-inverse* de la matrice A de type (m, n) la matrice A^+ dont les colonnes sont les pseudo-solutions des systèmes d'équations linéaires de la forme (6), où e_i sont les colonnes de la matrice unité d'ordre m.

Ceci étant, A^+ est composée de m colonnes à n éléments, c'est-à-dire est de même type (n, m) que la matrice ${}^tA$.

Il découle du théorème 1 que toute matrice admet une seule matrice pseudo-inverse.

Pour une matrice carrée régulière A, la pseudo-solution de chacun des systèmes (6) se confond avec la solution et par suite, la matrice pseudo-inverse coïncide avec la matrice inverse.

Dans un autre cas particulier, celui de la matrice nulle à m lignes et n colonnes, on peut conclure en appliquant le résultat obtenu dans l'exemple 4 ci-dessus que la matrice pseudo-inverse est la matrice nulle à n lignes et m colonnes.

PROPOSITION 5. *La pseudo-solution du système d'équations linéaires* (1) *peut être écrite sous la forme* $x_0 = A^+b$.

En effet, la matrice-colonne des termes constants b est une combinaison linéaire des colonnes de la matrice unité d'ordre m :

$$b = \beta_1 e_1 + \ldots + \beta_m e_m.$$

D'après la définition de la matrice pseudo-inverse et selon la proposition 4, la pseudo-solution x_0 est une combinaison linéaire des colonnes de la

matrice pseudo-inverse, dont les coefficients sont les mêmes :

$$x_0 = \beta_1 a_1^+ + \dots + \beta_m a_m^+.$$

Ce qui est équivalent à l'assertion qu'il fallait démontrer.

Notons que la proposition 5 est en général d'une importance théorique, de même que la règle de Cramer pour les matrices régulières. La recherche de la matrice pseudo-inverse n'est pas obligatoire pour le calcul de la pseudo-solution normale et exige de grands efforts.

Rappelons que la norme euclidienne $\|A\|_E$ de la matrice A est la racine carrée de la somme des carrés de ses éléments (comp. § 4, ch. XI). La matrice pseudo-inverse possède la propriété d'extrémum suivante.

PROPOSITION 6. *Pour toute matrice X à n lignes et m colonnes on a la relation*

$$\|AA^+ - E\|_E \leqslant \|AX - E\|_E.$$

De plus, si pour une matrice quelconque X différente de A^+ se réalise l'égalité, on a $\|A^+\|_E < \|X\|_E$.

DÉMONSTRATION. De par la définition, pour tout i la colonne a_i^+ de la matrice pseudo-inverse donne par substitution dans le système (6) un écart minimal. Aussi pour la i-ième colonne de la matrice X a-t-on

$$\|Aa_i^+ - e_i\| \leqslant \|Ax_i - e_i\|.$$

Si on aboutit à une égalité pour $x_i \neq a_i^+$, on a $\|a_i^+\| < \|x_i\|$. Notons que le carré de la norme euclidienne de la matrice est égal à la somme des carrés des normes de ses colonnes. Donc, en élevant au carré et en sommant les relations données pour tous les $i = 1, \dots, m$, on aboutit à la proposition nécessaire.

D'après la formule (5), il existe pour tout $i = 1, \dots, m$ une matrice-colonne z_i telle que la i-ième colonne de la matrice pseudo-inverse prend la forme $a_i^+ = {}^t\!Az_i$. Donc, il existe une matrice carrée Z d'ordre m telle que

$$A^+ = {}^t\!AZ. \tag{7}$$

Cette matrice est la matrice $Z = \|z_1, \dots, z_m\|$ composée des matrices-colonnes z_i.

La pseudo-solution A^+b du système $Ax = b$ vérifie le système normal associé et, par suite,

$${}^t\!AAA^+b = {}^t\!Ab.$$

Cette égalité a lieu pour toute matrice-colonne des termes constants b, si bien que les matrices dans le premier et dans le second membre de l'égalité sont égales. Donc

$${}^t\!AAA^+ = {}^t\!A. \tag{8}$$

Parfois il est utile de formuler ce résultat d'une façon analogue à (7) : il existe une matrice carrée Z_1 d'ordre n pour laquelle

$$Z_1 A^+ = {}^tA.$$

PROPOSITION 7. *La matrice X est une matrice pseudo-inverse de A si et seulement si*

$${}^tAAX = {}^tA$$

et il existe une matrice carrée Z telle que $X = {}^tAZ$.

La nécessité de ces conditions découle des formules (7) et (8). Démontrons la suffisance. A cet effet, remarquons que pour la i-ième colonne de la matrice unité il ressort de ${}^tAAX = {}^tA$ que

$${}^tAAXe_i = {}^tAe_i,$$

ce qui signifie que Xe_i, c'est-à-dire la i-ième colonne de X, vérifie le système normal associé au i-ième système (6). De plus, cette matrice-colonne satisfait à la condition (5), avec la i-ième colonne de la matrice Z :

$$Xe_i = {}^tAZe_i.$$

La proposition est démontrée.

En se servant de la proposition 7, on établit facilement que pour tout nombre α différent de zéro

$$(\alpha A)^+ = \alpha^{-1} A^+.$$

De la proposition 7 découle aussi la

PROPOSITION 8. *Si U est une matrice orthogonale d'ordre m, on a*

$$(UA)^+ = (A^+)({}^tU),$$

et pour une matrice orthogonale d'ordre n,

$$(AU)^+ = {}^tUA^+.$$

Démontrons la première égalité. Pour la matrice A on a selon (7) $(A^+){}^tU = {}^tAZ{}^tU$. D'où $(A^+){}^tU = {}^tA{}^tUUZ{}^tU = {}^t(UA)(UZ{}^tU)$. Cela signifie que la matrice $(A^+){}^tU$ est le produit de ${}^t(UA)$ par la matrice carrée $Z_1 = UZ{}^tU$. Ensuite, en vertu de l'égalité (8), on a pour la matrice A

$${}^tA\,{}^tUUA(A^+)\,{}^tU = {}^tA\,{}^tU.$$

Ainsi, on a démontré la première partie de la proposition. La seconde partie se démontre de façon analogue.

Remarquons que l'hypothèse sur l'orthogonalité de la matrice U est ici essentielle. Déjà pour la matrice carrée régulière S, on a

$$(SA)^+ \neq A^+S^{-1},$$

comme le montre l'exemple suivant. Soient

$$S = \begin{Vmatrix} 1 & 2 \\ 2 & 0 \end{Vmatrix}, \quad A = \begin{Vmatrix} 1 \\ 0 \end{Vmatrix} \quad \text{et} \quad B = SA = \begin{Vmatrix} 1 \\ 2 \end{Vmatrix}.$$

La matrice ${}'BB$ possède le seul élément 5, de sorte que la formule (8) donne $B^+ = \begin{Vmatrix} \frac{1}{5}, & \frac{2}{5} \end{Vmatrix}$. D'autre part, on voit aussitôt que

$$S^{-1} = \begin{Vmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{4} \end{Vmatrix}, \quad A^+ = \| 1, 0 \| \quad \text{et} \quad A^+S^{-1} = \begin{Vmatrix} 0, & \frac{1}{2} \end{Vmatrix}.$$

Ainsi, la formule connue d'inversion du produit de matrices ne peut être généralisée pour les matrices pseudo-inverses :

$$(AB)^+ \neq B^+A^+.$$

Le lecteur désireux de connaître les conditions nécessaires et suffisantes pour que l'égalité $(AB)^+ = B^+A^+$ soit réalisée, ainsi que l'expression dans le cas général de la matrice pseudo-inverse du produit, peut les trouver dans le livre d'Albert [1].

La proposition suivante permet de trouver la matrice pseudo-inverse dans deux cas particuliers importants.

PROPOSITION 9. *Si les colonnes de la matrice A sont linéairement indépendantes,*

$$A^+ = ({}'AA)^{-1}({}'A). \tag{9}$$

Si les lignes de la matrice A sont linéairement indépendantes,

$$A^+ = {}'A(A\,{}'A)^{-1}. \tag{10}$$

Démontrons la première assertion. Si les colonnes de A sont linéairement indépendantes, le système normal ${}'AAx = {}'Ab$ présente d'après la proposition 3 une solution unique pour toute matrice-colonne b. La matrice ${}'AA$ possède dans ce cas une inverse, et la solution est donc égale à $({}'AA)^{-1}({}'A)b$. En remplaçant b par les colonnes de la matrice unité d'ordre m, on constate que les colonnes de A^+ sont les colonnes de la matrice du second membre de (9).

Si les lignes de la matrice A sont linéairement indépendantes, le système $Ax = b$ est compatible pour tout b selon le théorème de Kronecker-Capelli et, par suite, la solution dont la norme est minimale est sa pseudo-solution normale. Cette solution est de la forme $x = {}'Az$ avec un certain z. Il nous faut donc rechercher la matrice-colonne z qui vérifie le système d'équations linéaires $A\,{}'Az = b$.

Dans le cas considéré, la matrice $A\,{}^tA$ possède une inverse et, par suite, $z = (A\,{}^tA)^{-1}b$, de sorte que la pseudo-solution normale du système $Ax = b$ est la matrice-colonne $x = {}^tA\,(A\,{}^tA)^{-1}b$. D'où, en remplaçant b par les colonnes de la matrice unité, on arrive à la formule (10).

Le calcul de la matrice pseudo-inverse de toute matrice non nulle A peut être ramené à deux cas décrits dans la proposition 9 si on décompose A en produit de facteurs appelé décomposition squelettique.

On appelle *décomposition squelettique* d'une matrice A de type (m, n) et de rang $r > 0$ la décomposition de la forme $A = BC$, où B et C sont des matrices de type (m, r) et (r, n) respectivement.

Vu que le rang du produit ne dépasse pas ceux des facteurs et que les rangs de B et C ne peuvent être supérieurs à r, ils sont égaux à r.

On peut obtenir la décomposition squelettique d'une matrice non nulle par exemple de la façon suivante. Considérons les colonnes de la matrice A qui renferment son mineur principal et composons à partir d'elles la matrice B. C'est une matrice à m lignes et r colonnes. D'après un théorème connu, chaque colonne de la matrice A est une combinaison linéaire des colonnes de la matrice B. Ecrivons les coefficients de ces combinaisons linéaires en colonnes et disposons-les dans l'ordre naturel, en obtenant par là même la matrice C. Cette matrice possède n colonnes à r éléments, c'est-à-dire qu'elle est de type (r, n).

On sait que la k-ième colonne du produit des matrices est une combinaison linéaire des colonnes du premier facteur dont les coefficients sont égaux aux éléments de la k-ième colonne du second facteur. Donc, $A = BC$ et on a obtenu l'une des décompositions squelettiques de la matrice A.

Sous l'angle de calcul numérique, ce procédé d'obtention de la décomposition squelettique est loin d'être parfait. On étudiera d'autres procédés dans le § 3.

On a vu que dans le cas général la formule $(BC)^+ = C^+B^+$ est fausse. Toutefois, on a la proposition suivante.

PROPOSITION 10. *Si $A = BC$ est une décomposition squelettique de la matrice A, sa matrice pseudo-inverse est*

$$A^+ = C^+B^+ = {}^tC(C\,{}^tC)^{-1}\,({}^tBB)^{-1}\,({}^tB).$$

DÉMONSTRATION. Vu que le rang de B et de C est r, les lignes de C et les colonnes de B sont linéairement indépendantes. En vertu de la proposition 9, $C^+ = {}^tC(C\,{}^tC)^{-1}$ et $B^+ = ({}^tBB)^{-1}({}^tB)$. Vérifions pour la matrice $R = {}^tC(C\,{}^tC)^{-1}({}^tBB)^{-1}({}^tB)$ la condition (8). En portant ${}^tAA = {}^tC\,{}^tBBC$, on a

$${}^tC\,{}^tBBC\,{}^tC\,(C\,{}^tC)^{-1}\,({}^tBB)^{-1}\,({}^tB) = {}^tC\,{}^tB,$$

ce qu'il fallait démontrer.

La condition (7) imposée à la matrice R signifie qu'il existe une matrice carrée Z pour laquelle $R = {}^tC{}^tBZ$. La condition sera vérifiée si l'on démontre l'existence d'une matrice carrée Z d'ordre m pour laquelle ${}^tBZ = (C{}^tC)^{-1}({}^tBB)^{-1}({}^tB)$. La dernière égalité est équivalente aux systèmes d'équations linéaires de la forme

$$ {}^tBz_i = s_i \quad (i = 1, \ldots, m) $$

pour les colonnes de la matrice Z, où le second membre du i-ième système est la i-ième colonne de la matrice $S = (C{}^tC)^{-1}({}^tBB)^{-1}({}^tB)$. (Il est aisé de vérifier que la matrice S est de type (r, m).) Les lignes de la matrice ${}^tB$, autrement dit les colonnes de la matrice B, sont linéairement indépendantes, et par suite, chacun des systèmes est compatible en vertu du théorème de Kronecker-Capelli.

Ainsi, les deux conditions (7) et (8) sont vraies pour la matrice R, si bien que la proposition 10 découle de la proposition 7.

La proposition 10 permet d'obtenir certaines propriétés des matrices pseudo-inverses. A savoir, on a l'égalité

$$ {}^t(A^+) = ({}^tA)^+. \tag{11} $$

En effet, les matrices ${}^tBB$ et $C\,{}^tC$ sont symétriques. Il en est de même de leurs inverses. Aussi, en transposant l'expression de A^+, aboutit-on à

$$ {}^t(A^+) = B({}^tBB)^{-1}\,(C\,{}^tC)^{-1}\,C. $$

Or ${}^tA = {}^tC\,{}^tB$ est une décomposition squelettique de ${}^tA$. En l'appliquant, on obtient exactement la même expression pour $({}^tA)^+$. La démonstration donnée ne convient pas à la matrice nulle, mais pour cette matrice la formule (11) est évidente. On achève ainsi la démonstration de la formule (11).

Ensuite, en utilisant toujours la même proposition 10, on obtient

$$ A^+A = {}^tC(C\,{}^tC)^{-1}C, \tag{12} $$

d'où on voit, en particulier, que la matrice A^+A est symétrique.

Admettons comme plus haut que $\mathscr{K}$ désigne l'ensemble des solutions du système $A\boldsymbol{z} = \boldsymbol{0}$, et $\mathscr{I}$ l'ensemble des matrices-colonnes de la forme ${}^tA\boldsymbol{b}$ pour tous les $\boldsymbol{b}$ possibles.

PROPOSITION 11. *Pour toute matrice-colonne $\boldsymbol{x}$, la matrice-colonne $A^+A\boldsymbol{x}$ est une projection orthogonale de $\boldsymbol{x}$ sur $\mathscr{I}$.*

DÉMONSTRATION. Si la matrice A est nulle, l'ensemble $\mathscr{I}$ n'est composé que du zéro, et la proposition est évidente. Pour une matrice A non nulle, on peut recourir à la formule (12). Représentons une matrice-colonne $\boldsymbol{x}$ à n éléments sous la forme $\boldsymbol{x}_0 + \boldsymbol{x}_1$, où $\boldsymbol{x}_0 \in \mathscr{I}$ et $\boldsymbol{x}_1 \in \mathscr{K}$. Alors $A^+A\boldsymbol{x}_1 = \boldsymbol{0}$ car

30—6442

$Ax_1 = \mathbf{0}$, si bien qu'on aboutit, en vertu des formules (5) et (12), à

$$A^+A(x_0 + x_1) = A^+Ax_0 = A^+A\,{}^tAz = {}^tC(C\,{}^tC)^{-1}C\,{}^tC\,{}^tBz = {}^tAz = x_0.$$

La proposition est démontrée.

Appliquons la formule (12) à la matrice ${}^tA$. On s'assure que la matrice $({}^tA)^+({}^tA)$ est symétrique et l'on obtient d'après la formule (11) que

$$({}^tA)^+({}^tA) = {}^t(({}^tA)^+({}^tA)) = AA^+.$$

Maintenant, en appliquant la proposition 11 à la matrice ${}^tA$, on obtient le

COROLLAIRE. *Pour toute matrice-colonne $y \in \mathscr{R}_m$, la matrice-colonne AA^+y est une projection orthogonale de y sur le sous-espace des matrices-colonnes de la forme Az pour tous les z possibles de $\mathscr{R}_n$.*

De la proposition 11 et de son corollaire on déduit que les matrices A^+A et AA^+ sont *idempotentes*, c'est-à-dire que

$$(A^+A)^2 = A^+A,$$
$$(AA^+)^2 = AA^+.$$

Chaque colonne de la matrice A^+ est une pseudo-solution normale du système (6) à matrice A et par suite, se trouve dans $\mathscr{I}$. En appliquant la proposition 11, on obtient $A^+Aa_i^+ = a_i^+$ pour toute matrice-colonne de A^+ et, par suite,

$$A^+AA^+ = A^+. \qquad (13)$$

D'une façon analogue, en se servant du corollaire de la proposition 11, on peut démontrer la formule

$$AA^+A = A. \qquad (14)$$

En utilisant la majoration du rang du produit de matrices dans les égalités (13) et (14), on obtient

$$\operatorname{Rg} A^+ = \operatorname{Rg} A^+AA^+ \leqslant \operatorname{Rg} A$$

et

$$\operatorname{Rg} A = \operatorname{Rg} AAA^+ \leqslant \operatorname{Rg} A^+,$$

d'où

$$\operatorname{Rg} A = \operatorname{Rg} A^+. \qquad (15)$$

PROPOSITION 12. *La matrice pseudo-inverse A^+ est la seule à vérifier les égalités*

$$AXA = A, \quad {}^t(AX) = AX$$

et

$$XAX = X, \quad {}^t(XA) = XA.$$

DÉMONSTRATION. On a vu plus haut que pour $X = A^+$ toutes ces égalités sont vérifiées. Démontrons qu'elles ne le sont que pour A^+.

Transposons la première égalité : ${}^tA\,{}^t(AX) = {}^tA$. D'où en vertu de la deuxième égalité on a ${}^tAAX = {}^tA$. Par ailleurs, en portant la quatrième égalité dans la troisième, on obtient

$$X = XAX = {}^t(XA)X = {}^tA\,{}^tXX,$$

ou $X = {}^tAZ$, avec $Z = {}^tXX$. Maintenant la proposition 12 découle de la proposition 7.

Utilisons les propriétés obtenues de la matrice pseudo-inverse à la recherche de la solution générale d'un système normal.

PROPOSITION 13. *La solution générale du système normal* (3) *associé au système* (1) *se définit par la formule*

$$x = A^+b + (E - A^+A)c,$$

où c est une matrice-colonne à n éléments. En particulier, si le système (1) *est compatible, cette formule définit sa solution générale.*

DÉMONSTRATION. Selon la proposition 5, la matrice-colonne A^+b est la pseudo-solution normale et par suite, est une solution particulière du système (3). Il reste à démontrer que la matrice-colonne $z = (E - A^+A)c$ est pour un c arbitraire la solution générale du système homogène normal ${}^tAAz = \mathbf{0}$. Démontrons-le.

D'abord, pour tout c

$${}^tAA[(E - A^+A)c] = {}^tAAc - {}^tAAA^+Ac = {}^tAAc - {}^tAAc = \mathbf{0}.$$

Cela signifie que z est la solution du système homogène normal.

Ensuite, pour toute solution z du système ${}^tAAz = \mathbf{0}$ il existe une matrice-colonne c pour laquelle

$$z = (E - A^+A)c.$$

On peut tout simplement poser $c = z$ car le système ${}^tAAz = \mathbf{0}$ est équivalent au système $Az = \mathbf{0}$ et, par suite,

$$(E - A^+A)z = z - A^+Az = z,$$

ce qui achève la démonstration de la proposition.

On peut obtenir l'interprétation géométrique de ce résultat en remarquant qu'en vertu de la proposition 11 la matrice-colonne $(E - A^+A)c$ est pour tout c une projection orthogonale de c sur le sous-espace $\mathscr{K}$ des solutions du système $Az = \mathbf{0}$.

§ 2. Application pseudo-inverse

Dans ce paragraphe, à la différence du § 1, on suivra le point de vue géométrique. On y définira et étudiera une application pseudo-inverse de l'application linéaire $\mathbf{A} : \mathscr{E}_n \to \tilde{\mathscr{E}}_m$, où $\mathscr{E}_n$ et $\tilde{\mathscr{E}}_m$ sont des espaces euclidiens de dimension n et m.

1. Définition. L'application $\mathbf{A}$ peut s'avérer non bijective pour deux raisons : ou bien $\operatorname{Ker} \mathbf{A} \neq o$ et l'on obtient alors deux vecteurs différents x_1 et x_2 dans $\mathscr{E}_n$ dont les images coïncident ; ou bien $\operatorname{Im} \mathbf{A}$ ne coïncide pas avec $\tilde{\mathscr{E}}_m$ et il existe alors dans $\tilde{\mathscr{E}}_m$ un vecteur qui n'a pas d'antécédent. Rappelons que pour une application adjointe $\mathbf{A}^* : \tilde{\mathscr{E}}_m \to \mathscr{E}_n$ est vérifiée l'égalité $\operatorname{Im} \mathbf{A}^* = (\operatorname{Ker} \mathbf{A})^{\perp}$. Aussi $\mathbf{A}$ est-elle une bijection si et seulement si sont réalisées deux conditions :

$$\operatorname{Im} \mathbf{A}^* = \mathscr{E}_n, \quad \operatorname{Im} \mathbf{A} = \tilde{\mathscr{E}}_m.$$

On peut transformer $\mathbf{A}$ en une application bijective si l'on se limite aux sous-espaces des espaces donnés. Plus précisément on a la proposition suivante.

PROPOSITION 1. *Soit* $\mathbf{A}_0$ *la restriction de* $\mathbf{A}$ *à* $\operatorname{Im} \mathbf{A}^*$. *Alors* $\mathbf{A}_0$ *applique* $\operatorname{Im} \mathbf{A}^*$ *sur* $\operatorname{Im} \mathbf{A}$ *de façon bijective.*

En effet, considérons un vecteur arbitraire $y \in \operatorname{Im} \mathbf{A}$. Il a un antécédent x dans $\mathscr{E}_n$. Décomposons x en somme $x_0 + x_1$, où $x_0 \in \operatorname{Im} \mathbf{A}^*$ et $x_1 \in \operatorname{Ker} \mathbf{A}$. Alors $\mathbf{A}(x) = \mathbf{A}(x_0) + \mathbf{A}(x_1) = \mathbf{A}(x_0)$. Donc, l'antécédent de y est le vecteur x_0 de $\operatorname{Im} \mathbf{A}^*$.

Montrons que cet antécédent est unique dans $\operatorname{Im} \mathbf{A}^*$. En effet, soient $x_0, x_0' \in \operatorname{Im} \mathbf{A}^*$ et $\mathbf{A}(x_0) = \mathbf{A}(x_0')$. Cela signifie que $x_0 - x_0' \in \operatorname{Im} \mathbf{A}^*$ et $x_0 - x_0' \in \operatorname{Ker} \mathbf{A}$, d'où $x_0 - x_0' = o$.

L'application $\mathbf{A}_0$ introduite dans la proposition 1, comme toute application bijective, possède une application inverse $\mathbf{A}_0^{-1}$ définie sur le sous-espace $\operatorname{Im} \mathbf{A} \subseteq \tilde{\mathscr{E}}_m$. On veut définir une application linéaire $\mathbf{A}^+ : \tilde{\mathscr{E}}_m \to \mathscr{E}_n$ pour laquelle $\mathbf{A}_0^{-1}$ est une restriction à $\operatorname{Im} \mathbf{A}$. A cet effet, introduisons une application $\mathbf{P}$ qui associe à chaque vecteur $y \in \tilde{\mathscr{E}}_m$ la projection orthogonale de y sur $\operatorname{Im} \mathbf{A}$.

DÉFINITION. L'application $\mathbf{A}_0^{-1}\mathbf{P} : \tilde{\mathscr{E}}_m \to \mathscr{E}_n$ est dite *pseudo-inverse* de $\mathbf{A}$ et est notée $\mathbf{A}^+$.

PROPOSITION 2. *Quelles que soient les bases orthonormées e dans* $\mathscr{E}_n$ *et f dans* $\tilde{\mathscr{E}}_m$, *la matrice de l'application* $\mathbf{A}^+$ *par rapport aux bases f et e est pseudo-inverse de la matrice de l'application* $\mathbf{A}$ *par rapport aux bases e et f.*

DÉMONSTRATION. Vérifions d'abord que les applications $\mathbf{A}$, $\mathbf{A}^*$ et $\mathbf{A}^+$ sont liées par la relation

$$\mathbf{A}^* \mathbf{A} \mathbf{A}^+ = \mathbf{A}^*.$$

Considérons pour cela un vecteur $y \in \tilde{\mathscr{E}}_m$ et notons y' sa projection orthogonale sur Im **A**. L'image de y par l'application $\mathbf{A}^*\mathbf{A}\mathbf{A}^+$ est $\mathbf{A}^*\mathbf{A}\mathbf{A}_0^{-1}\mathbf{P}(y) = \mathbf{A}^*\mathbf{A}\mathbf{A}_0^{-1}(y') = \mathbf{A}^*(y')$.

D'autre part, $y - y' \in (\mathrm{Im}\,\mathbf{A})^{\perp} = \mathrm{Ker}\,\mathbf{A}^*$, et par suite, $\mathbf{A}^*(y) = \mathbf{A}^*(y')$. L'égalité est donc vérifiée. Si les bases sont orthonormées et l'application **A** a pour matrice A, ${}^tA$ est alors la matrice de l'application $\mathbf{A}^*$. Donc, la matrice de l'application $\mathbf{A}^+$ satisfait à la condition (8) du § 1.

Etablissons l'égalité (7) du § 1. D'après la définition de l'application $\mathbf{A}^+$ on a $\mathrm{Im}\,\mathbf{A}^+ \subseteq \mathrm{Im}\,\mathbf{A}^*$. Aussi pour chacun des vecteurs de base f_i ($i = 1, \ldots, m$) existe-t-il un vecteur z_i tel que $\mathbf{A}^+(f_i) = \mathbf{A}^*(z_i)$. Cela signifie en coordonnées que la i-ième colonne de la matrice associée à l'application $\mathbf{A}^+$ est de la forme ${}^tA\mathfrak{z}_i$ pour une matrice-colonne $\mathfrak{z}_i$ appropriée. Ceci est équivalent à l'égalité qu'il fallait vérifier.

On s'est assuré que dans un couple de bases orthonormées la matrice de l'application pseudo-inverse vérifie les égalités (8) et (7) du § 1. Pour achever la démonstration, il ne reste qu'à se référer à la proposition 7 du § 1.

Remarquons que l'inclusion $\mathrm{Im}\,\mathbf{A}^+ \subseteq \mathrm{Im}\,\mathbf{A}^*$ écrite lors de la démonstration de la proposition 2 donne avec l'égalité des rangs (15) du § 1

$$\mathrm{Im}\,\mathbf{A}^+ = \mathrm{Im}\,\mathbf{A}^*. \tag{1}$$

La proposition 2 permet de rapporter les résultats sur les matrices pseudo-inverses obtenus au § 1 aux applications pseudo-inverses. Notons, par exemple, que

$$(\mathbf{A}^+)^* = (\mathbf{A}^*)^+. \tag{2}$$

Mais on ne s'arrêtera pas systématiquement sur ce fait. Notre objectif immédiat est de simplifier la matrice de l'application $\mathbf{A}^+$ par un choix convenable de bases.

2. Application pseudo-inverse en bases singulières. Il existe pour l'application **A** des bases singulières $\boldsymbol{e}$ dans $\mathscr{E}_n$ et $\boldsymbol{f}$ dans $\tilde{\mathscr{E}}_m$, telles que la matrice de l'application **A** par rapport à ces bases se met sous la forme indiquée au théorème 1 du § 1, ch. XI :

$$D = \left\| \begin{array}{ccc|c} \alpha_1 & & & O \\ & \ddots & & \\ & & \alpha_r & \\ \hline & O & & O \end{array} \right\|. \tag{3}$$

Les seuls éléments non nuls de cette matrice sont de la forme $d_{ii} = \alpha_i$ pour $i \leqslant r = \mathrm{Rg}\,\mathbf{A}$.

Les matrices de cette forme généralisent les matrices diagonales carrées. La matrice rectangulaire d'éléments $d_{ij} = 0$ pour $i \neq j$ sera appelée *matrice diagonale*.

Les éléments non nuls de la matrice (3) seront appelés nombres singuliers non nuls de l'application **A**.

Vu que les bases singulières sont orthonormées, l'application pseudo-inverse $\mathbf{A}^+$ de l'application **A** se définit dans les bases singulières par la matrice D^+.

Cherchons la matrice pseudo-inverse de la matrice diagonale. Sa i-ième colonne est pour tout $i = 1, \ldots, m$ la pseudo-solution normale du système d'équations linéaires $D\boldsymbol{\xi} = \boldsymbol{\varepsilon}_i$, où $\boldsymbol{\varepsilon}_i$ est la i-ième colonne de la matrice unité d'ordre m. Le système d'équations normal associé à ce système, soit :

$$ {}^tDD\boldsymbol{\xi} = {}^tD\boldsymbol{\varepsilon}_i, $$

peut être écrit d'une façon plus détaillée :

$$ \begin{aligned} &\alpha_j^2 \xi^j = \alpha_j \cdot 0 && \text{pour} \quad j \neq i, \quad j \leqslant r, \\ &\alpha_i^2 \xi^i = \alpha_i && \text{pour} \quad i \leqslant r, \\ &0 \cdot \xi^k = 0 && \text{pour} \quad k > r. \end{aligned} $$

Il va de soi que la matrice-colonne $\boldsymbol{\xi}$ sera solution de ce système si $\xi^i = \alpha_i^{-1}$ pour $i \leqslant r$, et $\xi^j = 0$ pour $j \neq i, j \leqslant r$, quels que soient les éléments ξ^k pour $k > r$. La solution possédera la norme minimale si tous ces éléments sont nuls, c'est-à-dire si $\xi^k = 0$ pour $k > r$. Ainsi donc, la i-ième colonne de la matrice D^+ est égale à ${}^t\|0 \ldots \alpha_i^{-1} 0 \ldots 0\|$ pour $i \leqslant r$ et à $\mathbf{0}$ pour $i > r$. La matrice D^+ composée de ces colonnes est diagonale de type (n, m), avec les nombres $\alpha_1^{-1}, \ldots, \alpha_r^{-1}$ sur la « diagonale » :

$$ D^+ = \left\| \begin{array}{ccc|c} \alpha_1^{-1} & & & O \\ & \ddots & & \\ & & \alpha_r^{-1} & \\ \hline & O & & O \end{array} \right\|. \tag{4} $$

Enumérons quelques conséquences immédiates du résultat obtenu. Vu que D^+ est une matrice diagonale, sa matrice pseudo-inverse se construit suivant la même règle et l'on constate que $(D^+)^+ = D$. Il s'ensuit que pour une application arbitraire

$$ (\mathbf{A}^+)^+ = \mathbf{A}. \tag{5} $$

En écrivant la relation analogue entre les matrices rapportées à un couple de bases orthonormées, on obtient la propriété suivante de l'opération de passage à la matrice pseudo-inverse :

$$ (A^+)^+ = A. \tag{6} $$

La comparaison des matrices D^+ et ${}^tD$ montre que l'égalité $\mathbf{A}^+ = \mathbf{A}^*$ est vérifiée si et seulement si les nombres singuliers non nuls de **A** sont égaux à 1. Cette propriété généralise la propriété de la transformation orthogonale pour laquelle $\mathbf{A}^{-1} = \mathbf{A}^*$ et tous les nombres singuliers sont égaux à 1.

Désignons par e et f respectivement la première et la seconde base singulière de l'application **A**. Soit D la matrice de **A** par rapport aux bases e et f. D^+ est alors la matrice de $\mathbf{A}^+$ par rapport aux bases f et e, et ${}^t(D^+)D^+$ la matrice de l'application $(\mathbf{A}^+)^*\mathbf{A}^+$ dans la base f.

On voit aussitôt que

$$^t(D^+)\,D^+ = \operatorname{diag}\,(\alpha_1^{-2}, \ldots, \alpha_r^{-2}, 0, \ldots, 0)$$

est une matrice diagonale carrée d'ordre m. Cela signifie que la seconde base singulière f de l'application **A** est composée des vecteurs propres de l'application ${}^t(\mathbf{A}^+)\mathbf{A}^*$. Elle ne peut donc différer de la première base singulière de l'application $\mathbf{A}^+$ que par l'ordre des vecteurs. (Rappelons que les vecteurs de la base singulière sont ordonnés de manière que la suite des nombres singuliers soit décroissante.)

Par ailleurs, les images des vecteurs de la base f par l'application $\mathbf{A}^+$ sont des vecteurs dont les colonnes de coordonnées s'identifient aux colonnes de la matrice D^+. On les obtient en multipliant certains vecteurs de la base e par les facteurs α_i^{-1} ou 0. La base e est une base orthonormée dans $\mathscr{E}_n$ qui contient des vecteurs normés non nuls $\mathbf{A}^+(f_i)$. Aussi la base e ne diffère-t-elle de la seconde base singulière de l'application $\mathbf{A}^+$ que par l'ordre des vecteurs. On obtient donc la

Proposition 3. *La première et la seconde base singulière de l'application* $\mathbf{A}^+$ *ne diffèrent respectivement de la seconde et de la première base singulière de l'application* **A** *que par l'ordre des vecteurs. Si* $\alpha_1, \ldots, \alpha_r$ *sont des nombres singuliers non nuls de* **A**, *alors* $\alpha_r^{-1}, \ldots, \alpha_1^{-1}$ *sont ceux de* $\mathbf{A}^+$.

Pour toute application, le noyau est l'enveloppe linéaire des vecteurs de la première base singulière qui correspondent aux nombres singuliers nuls. Aussi bien pour $\mathbf{A}^+$ que pour $\mathbf{A}^*$, ce sont les vecteurs $f_{r+1}, \ldots, f_m$. Donc,

$$\operatorname{Ker}\mathbf{A}^+ = \operatorname{Ker}\mathbf{A}^*. \tag{7}$$

Le passage aux bases singulières permet de vérifier facilement diverses identités se rapportant aux matrices pseudo-inverses. Démontrons par exemple les égalités qui généralisent les formules (9) et (10) du § 1 :

$$A^+ = ({}^tAA)^+({}^tA) \tag{8}$$

et

$$A^+ = {}^tA(A\,{}^tA)^+. \tag{9}$$

Il ressort de la proposition 2 que l'égalité (8) est équivalente à la décomposition suivante de l'application pseudo-inverse :

$$\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^+\mathbf{A}^*.$$

Cette égalité est invariante et, par suite, il suffit de la vérifier pour un cou-

ple quelconque de bases. Vérifions-la pour les bases singulières de l'application **A**.

Dans ce cas, la matrice D de l'application **A** est de la forme (3). En appliquant la formule (4) à la matrice diagonale $'DD$, on obtient

$$('DD)^{+} = \operatorname{diag}(\alpha_1^{-2}, \ldots, \alpha_r^{-2}, 0, \ldots, 0).$$

On voit aussitôt que le produit $('DD)^{+}('D)$ se confond avec D^{+}. Ce qu'il fallait justement. La formule (9) se démontre de façon analogue.

3. Pseudo-inversion par passage à la limite. Désignons par $\mathscr{H}$ l'ensemble de toutes les applications linéaires de $\tilde{\mathscr{E}}_m$ dans $\mathscr{E}_n$. Les opérations d'addition et de multiplication par un nombre dans $\mathscr{H}$ ont été introduites dans le chapitre XI. On voit aisément que l'ensemble $\mathscr{H}$ muni de ces opérations est un espace vectoriel et que le choix de bases dans $\tilde{\mathscr{E}}_m$ et $\mathscr{E}_n$ établit un isomorphisme de $\mathscr{H}$ sur l'espace $\mathscr{M}_{n,m}$ des matrices à n lignes et m colonnes.

Supposons qu'à chaque nombre λ de l'intervalle $]\alpha, \beta[$ est associée une application $\mathbf{B}(\lambda) \in \mathscr{H}$. On dira dans ce cas que sur $]\alpha, \beta[$ est définie une fonction à valeurs dans $\mathscr{H}$.

Pour définir la notion de limite pour une fonction à valeurs dans $\mathscr{H}$, introduisons dans $\mathscr{H}$ une norme. On peut le faire par exemple ainsi. Rapportons les espaces $\tilde{\mathscr{E}}_m$ et $\mathscr{E}_n$ à deux bases orthonormées et considérons la matrice B de l'application $\mathbf{B} \in \mathscr{H}$. Sa norme spectrale ne varie pas lorsqu'on multiplie B à gauche et à droite par des matrices orthogonales, c'est-à-dire que $\|B\| = \|UBV\|$ pour des matrices orthogonales quelconques U et V d'ordre n et m respectivement. Aussi la norme spectrale de la matrice de l'application **B** est-elle indépendante du choix de bases orthonormées et peut être prise pour norme de l'application **B**. On la notera $\|\mathbf{B}\|$ en donnant la définition suivante.

L'application **B** est appelée *limite* de la fonction $\mathbf{B}(\lambda)$ pour $\lambda \to \mu$ si pour tout $\varepsilon > 0$ il existe un nombre δ tel que $0 < |\lambda - \mu| < \varepsilon$ entraîne $\|\mathbf{B}(\lambda) - \mathbf{B}\| < \varepsilon$.

On voit aussitôt que pour tout couple de bases orthonormées la relation

$$\lim_{\lambda \to \mu} \mathbf{B}(\lambda) = \mathbf{B}$$

entraîne une relation analogue pour les matrices des applications considérées :

$$\lim_{\lambda \to \mu} B(\lambda) = B.$$

Revenons maintenant à l'étude de l'application $\mathbf{A} : \mathscr{E}_n \to \tilde{\mathscr{E}}_m$ et de sa pseudo-inverse et considérons une transformation de l'espace $\tilde{\mathscr{E}}_m$ définie par l'expression $\mathbf{AA}^* + \alpha\tilde{\mathbf{E}}$, où $\tilde{\mathbf{E}}$ est la transformation indentique de $\tilde{\mathscr{E}}_m$.

Démontrons que pour $\alpha > 0$ cette transformation possède une inverse. Pour ne pas oublier le fait que α est strictement positif, désignons α par λ^2 et montrons que la matrice de la transformation $\mathbf{AA^*} + \lambda^2 \bar{\mathbf{E}}$ est régulière dans toute base orthonormée, et qu'elle est de plus définie positive. En effet, pour toute matrice-colonne ξ non nulle, il vient

$$ {}^t\xi (A\,{}^tA + \lambda^2 E)\xi = {}^t({}^tA\xi)({}^tA\xi) + \lambda^2({}^t\xi)\xi > 0. $$

Ainsi, pour tout nombre $\lambda \neq 0$ est définie une application $\mathbf{B}(\lambda) : \tilde{\mathscr{E}}_m \to \mathscr{E}_n$ suivant la formule

$$ \mathbf{B}(\lambda) = \mathbf{A^*}(\mathbf{AA^*} + \lambda^2 \bar{\mathbf{E}})^{-1}. $$

D'une façon analogue, on peut définir une application

$$ (\mathbf{A^*A} + \lambda^2 \mathbf{E})^{-1} \mathbf{A^*} $$

et démontrer l'existence de la transformation inverse de $\mathbf{A^*A} + \lambda^2 \mathbf{E}$ dans l'espace $\mathscr{E}_n$ comme on vient de le faire pour $\mathbf{AA^*} + \lambda^2 \bar{\mathbf{E}}$.

PROPOSITION 4. *On a les relations*

$$ \lim_{\lambda \to 0} \mathbf{A^*}(\mathbf{AA^*} + \lambda^2 \bar{\mathbf{E}})^{-1} = \mathbf{A}^+ $$

et

$$ \lim_{\lambda \to 0} (\mathbf{A^*A} + \lambda^2 \mathbf{E})^{-1} \mathbf{A^*} = \mathbf{A}^+. $$

Les deux formules se démontrent de façon identique. Démontrons donc la première. Pour cela, écrivons la matrice de l'application, qui se trouve sous le signe limite, par rapport aux bases singulières de l'application $\mathbf{A}$. Si la matrice A rapportée à ces bases est notée D, la matrice cherchée est ${}^tD(D\,{}^tD + \lambda^2 E)^{-1}$. En tenant compte de la forme (3) de la matrice D, on remarque que $C = (D\,{}^tD + \lambda^2 E)^{-1}$ est une matrice diagonale à éléments diagonaux

$$ (\alpha_1^2 + \lambda^2)^{-1}, \ldots, (\alpha_r^2 + \lambda^2)^{-1}, \lambda^{-2}, \ldots, \lambda^{-2}. $$

La matrice ${}^tDC$ est matrice diagonale de même type (n, m) que ${}^tD$ et présente sur la diagonale r éléments non nuls

$$ \alpha_1 (\alpha_1^2 + \lambda^2)^{-1}, \ldots, \alpha_r (\alpha_r^2 + \lambda^2)^{-1} $$

(les éléments de la matrice C, égaux à λ^{-2}, ont été multipliés par 0). Quand $\lambda \to 0$, la matrice ${}^tDC$ tend au sens de convergence en éléments vers la matrice D^+ définie par la fourmule (4). Elle doit tendre aussi vers cette matrice en norme spectrale. D'où il découle immédiatement la relation nécessaire.

La proposition 4 entraîne les expressions suivantes pour la matrice

pseudo-inverse :

$$\lim_{\lambda \to 0} {}^t A (A {}^t A + \lambda^2 E)^{-1} = A^+, \tag{10}$$

$$\lim_{\lambda \to 0} ({}^t A A + \lambda^2 E)^{-1} ({}^t A) = A^+. \tag{11}$$

Considérons le système d'équations linéaires à matrice régulière

$$({}^t A A + \lambda^2 E) \xi = {}^t A b. \tag{12}$$

Désignons sa solution par ξ_λ. Alors

$$\xi_\lambda = ({}^t A A + \lambda^2 E)^{-1} ({}^t A) b \tag{13}$$

et la formule (11) montre qu'est vraie la

PROPOSITION 5. *La solution ξ_λ du système* (12) *tend pour $\lambda \to 0$ vers la pseudo-solution normale du système $A\xi = b$.*

Cette proposition est d'une grande importance théorique. On sait que la pseudo-solution normale du système d'équations linéaires n'est pas une fonction continue de la matrice du système. La proposition 5 montre que le système donné peut être inclus dans une famille de systèmes de paramètre λ, de manière que la solution du système dépend continûment de ce paramètre. Ce résultat a été obtenu dans un contexte plus général de la théorie des fonctionnelles de régularisation pour les problèmes mal posés (voir Tykhonov et Arsénine [38]). La théorie mentionnée se rapporte en général aux équations dans les espaces de dimension infinie (par exemple, aux équations aux dérivées partielles).

Dans le cas des espaces de dimension finie, la nécessité d'introduire les fonctionnelles de régularisation ne se présente pas directement. Toutefois, notons que pour des systèmes d'équations algébriques linéaires, le rôle de fonctionnelle de régularisation peut être joué par la fonction

$$f_\lambda(\xi, b, A) = \|b - A\xi\|^2 + \lambda^2 \|\xi\|^2$$

définie sur l'espace arithmétique $\mathscr{R}_n$. Cherchons la valeur de ξ pour laquelle elle présente son minimum. La fonction f_λ peut être écrite d'une autre façon :

$$f_\lambda(\xi, b, A) = {}^t(b - A\xi)(b - A\xi) + \lambda^2({}^t\xi)\xi.$$

Cherchons la différentielle en ξ de cette expression. Il vient

$$\mathrm{d}f_\lambda(\xi, b, A) = -2\mathrm{d}\,{}^t\xi\,{}^t A b + 2\mathrm{d}\,{}^t\xi\,{}^t A A \xi + 2\lambda^2\,\mathrm{d}\,{}^t\xi\,\xi.$$

La différentielle s'annule pour les matrices-colonnes ξ vérifiant le système d'équations qui coïncide avec (12). On a vu plus haut que le déterminant de la matrice du système est différent de zéro et que le système possède une solution unique (13) pour tous b, A et $\lambda \neq 0$.

Désignons cette solution par ξ_λ, et soit $f_\lambda(\xi_\lambda, b, A) = \zeta$. Si $\|\xi\| > \sqrt{\zeta}/\lambda$, on obtient $f_\lambda(\xi, b, A) > \zeta$. Il s'ensuit que sur la sphère de rayon $\sqrt{\zeta}/\lambda + 1$ et en dehors d'elle, f_λ prend des valeurs strictement supérieures à ζ. Si ξ_λ ne tombe pas à l'intérieur de la sphère, augmentons son rayon jusqu'à $\|\xi_\lambda\| + 1$. On obtient ainsi une sphère contenant ξ_λ et telle que pour tous les points situés sur la sphère et en dehors d'elle on a $f_\lambda(\xi, b, A) > \zeta$. Comme la fonction f_λ est continue et possède un point stationnaire unique à l'intérieur de la sphère, ce point est le point de son minimum. Les raisonnements donnés montrent que c'est un minimum absolu.

La proposition 5 veut dire en fait que pour $\lambda \to 0$, le point où la fonctionnelle de régularisation atteint le minimum tend vers la pseudo-solution du système $A\xi = b$.

§ 3. Méthodes de calcul

1. Recherche de la pseudo-solution par décomposition singulière. La plus grande partie du chapitre XIII a été consacrée aux systèmes d'équations linéaires à matrices carrées régulières, plus précisément aux systèmes dont les matrices ne sont pas quasi singulières. Toutefois, les problèmes réels ne garantissent pas en général des matrices régulières et, à plus forte raison, des matrices suffisamment bien conditionnées. (Font exception certains systèmes d'équations linéaires dont les matrices sont de forme spéciale.) Pour vérifier si une matrice est bien conditionnée, il faut, comme on l'a vu, autant d'efforts que pour résoudre le système.

Aussi, au cas où il est bien probable que la matrice du système

$$Ax = b \tag{1}$$

est quasi singulière, peut-il s'avérer préférable de rechercher aussitôt la pseudo-solution de ce système. On minimise la norme de l'écart. Si le minimum s'avère suffisamment petit, on considère qu'on a trouvé une solution, sinon on pose qu'on a une pseudo-solution et que la solution n'existe pas.

Il faut toutefois se rappeler que si la solution du système normal

$${}^tAAx = {}^tAb \tag{2}$$

n'est pas unique, la pseudo-solution normale est celle des solutions de (2) qui a la plus petite norme. Aussi peut-il s'avérer nécessaire de rechercher une autre solution du système normal.

Une grande partie de recommandations données pour résoudre les systèmes d'équations linéaires de forme générale prévoit de rechercher la décomposition singulière de la matrice du système ou de procéder à des opérations se réduisant en fait à cette dernière recherche. Si dans le système (1) on porte, au lieu de la matrice A, sa décomposition singulière, on

obtient le système

$$ {}^{t}UDVx = b, $$

ou

$$ Dy = \bar{b}, \tag{3} $$

avec

$$ b = Vx \tag{4} $$

et

$$ \bar{b} = Ub. $$

Le système normal (2) peut de même être reduit à la forme

$$ D^2 y = D\bar{b}. \tag{5} $$

En tenant compte de la forme de la matrice D, on trouve

$$ Dy = {}^{t}\| \alpha_1 y^1, \ldots, \alpha_r y^r, 0, \ldots 0 \|, $$

où $\alpha_1, \ldots, \alpha_r$ sont les nombres singuliers non nuls de la matrice A. Il est évident que pour la compatibilité du système (3) il faut et il suffit que

$$ \bar{b}^{r+1} = \ldots = \bar{b}^n = 0. $$

A cette condition, la matrice-colonne

$$ {}^{t}\| \alpha_1^{-1} \bar{b}^1, \ldots, \alpha_r^{-1} \bar{b}^r, y^{r+1}, \ldots, y^n \| \tag{6} $$

est sa solution quels que soient $y^{r+1}, \ldots, y^n$. Mais indépendamment de la compatibilité du système (3), la matrice-colonne (6) est la solution générale du système normal (5). La pseudo-solution normale, c'est-à-dire la matrice-colonne de norme minimale parmi les matrices-colonnes de la forme (6), est alors égale à :

$$ y_0 = {}^{t}\| \alpha_1^{-1} \bar{b}^1, \ldots, \alpha_r^{-1} \bar{b}^r, 0, \ldots, 0 \|. \tag{7} $$

Ainsi donc, la solution générale du système (5) (qui coïncide avec la solution générale du système (3) si ce dernier est compatible) est de la forme

$$ y = y_0 + c_1 e_{r+1} + \ldots + c_{n-r} e_n = y_0 + E_n^{n-r} c, $$

où c est une matrice-colonne quelconque à $n - r$ éléments, et E_n^{n-r} une matrice formée à partir des $n - r$ dernières colonnes $e_{r+1}, \ldots, e_n$ de la matrice unité d'ordre n.

Utilisons la formule (4) pour passer à la solution du système (1). On peut écrire

$$ x = {}^{t}Vy = {}^{t}Vy_0 + {}^{t}VE_n^{n-r} c. $$

Notons que la multiplication par une matrice orthogonale ${}^tV$ ne modifie pas la norme de la matrice-colonne, de sorte que la matrice-colonne $x_0 = {}^tVy_0$ est de norme minimale et par suite, est la pseudo-solution normale du système (1). De plus, il n'est pas difficile d'établir que la matrice ${}^tV' = {}^tVE_n^{n-r}$ est formée des $n - r$ dernières colonnes de la matrice ${}^tV$. Ainsi, la solution générale du système (2) peut être écrite par la formule

$$x = x_0 + {}^tV'c,$$

où x_0 est la pseudo-solution normale égale à ${}^tVy_0$, et c est une matrice-colonne quelconque à $n - r$ éléments.

Ainsi donc, si la décomposition singulière de la matrice du système (1) est connue, il devient facile d'obtenir la pseudo-solution normale de ce système, ainsi que sa solution générale si le système est compatible.

Il faut remarquer de plus que la recherche de la décomposition singulière est un processus beaucoup plus laborieux que la résolution du système, par exemple par la QR-décomposition de la matrice. L'obtention de la décomposition singulière sera étudiée au point 8 du présent paragraphe.

2. Utilisation de la régularisation. Une autre approche à la recherche de la pseudo-solution d'un système d'équations linéaires s'appuie sur la régularisation de sa matrice ou, plus simplement, sur l'utilisation de la proposition 5 du § 2. A savoir, pour approximation de la pseudo-solution du système (1), on prend la solution du système d'équations linéaires

$$({}^tAA + \lambda^2 E)x = {}^tAb, \tag{8}$$

avec une valeur convenable du paramètre λ. Dans la proposition 5 du § 2 on a montré que, si les données initiales sont précises et que les calculs soient réalisés exactement, la solution x_λ de ce système tend pour $\lambda \to 0$ vers la pseudo-solution x_0 du système (1).

Or pour des λ petits, la matrice du système devient en général mal conditionnée. Par conséquent, si les données initiales et les calculs sont entachés d'erreurs, la valeur calculée $\tilde{x}_\lambda$ de la solution x_λ peut différer fortement de celle de x_λ.

Etudions l'influence des perturbations de la matrice A et de la matrice-colonne des termes constants b sur la solution du système (8). Soit donné le système perturbé

$$\tilde{A}x = \tilde{b},$$

où la matrice $\tilde{A}$ et la matrice-colonne des termes constants $\tilde{b}$ vérifient les inégalités

$$\|\tilde{A} - A\| < \delta, \quad \|\tilde{b} - b\| < \delta \tag{9}$$

(on considère la norme spectrale de la matrice).

Soient x_λ la solution du système (8) et $\tilde{x}_\lambda$ la solution du système

$$({}^t\tilde{A}\tilde{A} + \lambda^2 E)\tilde{x}_\lambda = {}^t\tilde{A}\tilde{b}.$$

Estimons la norme de la différence $\tilde{x}_\lambda - x_\lambda$. De l'égalité

$$({}^t\tilde{A}\tilde{A} + \lambda^2 E)\tilde{x}_\lambda - ({}^tAA + \lambda^2 E)x_\lambda = {}^t\tilde{A}\tilde{b} - {}^tAb,$$

par des transformations simples on obtient

$$({}^t\tilde{A}\tilde{A} + \lambda^2 E)(\tilde{x}_\lambda - x_\lambda) = \\ = ({}^tAA - {}^t\tilde{A}\tilde{A})x_\lambda + ({}^t\tilde{A} - {}^tA)\tilde{b} + {}^tA(\tilde{b} - b),$$

ou

$$\tilde{x}_\lambda - x_\lambda = ({}^t\tilde{A}\tilde{A} + \lambda^2 E)^{-1}[({}^tAA - {}^t\tilde{A}\tilde{A})x_\lambda + \\ + ({}^t\tilde{A} - {}^tA)(\tilde{b} - b) + ({}^t\tilde{A} - {}^tA)b + {}^tA(\tilde{b} - b)]. \quad (10)$$

Pour tout x de norme égale à 1, il vient

$${}^tx({}^t\tilde{A}\tilde{A} + \lambda^2 E)x = {}^t(\tilde{A}x)\tilde{A}x + \lambda^2({}^tx)x \geqslant \lambda^2.$$

Il s'ensuit en vertu de la proposition 4 du § 2, ch. XI, que les nombres caractéristiques de la matrice ${}^t\tilde{A}\tilde{A} + \lambda^2 E$ sont supérieurs à λ^2 et que les nombres caractéristiques de son inverse sont inférieurs à λ^{-2}. Etant donné que pour une matrice symétrique les nombres singuliers sont égaux aux modules des nombres caractéristiques, le plus grand nombre singulier de $({}^t\tilde{A}\tilde{A} + \lambda^2 E)^{-1}$ ne dépasse pas λ^{-2}. Donc,

$$\| {}^t\tilde{A}\tilde{A} + \lambda^2 E \|^{-1} \leqslant \lambda^{-2}.$$

Pour la norme spectrale de la matrice ${}^tAA - {}^t\tilde{A}\tilde{A}$, on a

$$\| {}^tAA - {}^t\tilde{A}\tilde{A} \| \leqslant \| {}^tAA - {}^tA\tilde{A} \| + \| {}^tA\tilde{A} - {}^t\tilde{A}\tilde{A} \| \leqslant \\ \leqslant \| A - \tilde{A} \| (\| {}^tA \| + \| \tilde{A} \|) \leqslant \| A - \tilde{A} \| (2\|A\| + \| \tilde{A} - A \|) \leqslant \delta c + \delta^2.$$

On s'appuie ici sur le fait que la norme spectrale ne varie pas par transposition, ce qui résulte des propositions 6 et 14 du § 1, ch. XI. Notons que le nombre $c = 2\|A\|$ n'est défini que par la matrice A.

En se servant des estimations pour les normes des matrices figurant dans l'égalité (10), on obtient

$$\| \tilde{x}_\lambda - x_\lambda \| \leqslant \lambda^{-2}[(\delta c + \delta^2)\|x_\lambda\| + \delta^2 + \delta\|b\| + \|A\|\delta].$$

Etudions $x_0 = A^+b$ qui est la pseudo-solution du système (1). D'après la proposition 5 du § 2, $x_\lambda \to x_0$ pour $\lambda \to 0$. Aussi, pour des λ proches de zéro, la norme de x_λ est-elle bornée. D'où

$$\| \tilde{x}_\lambda - x_\lambda \| \leqslant \frac{\delta}{\lambda^2} p(\delta, \lambda), \quad (11)$$

où p est une fonction bornée au voisinage de zéro, qui ne dépend pas de $\tilde{A}$ et $\tilde{b}$.

Majorons la norme de la différence $x_0 - x_\lambda$:

$$\|x_0 - x_\lambda\| \leqslant \|A^+ - ({}^tAA + \lambda^2 E)^{-1}({}^tA)\| \cdot \|b\|.$$

Pour calculer la norme de la matrice $B = A^+ - ({}^tAA + \lambda^2 E)^{-1}({}^tA)$, considérons des matrices orthogonales U et V telles que $D = {}^tUAV$ soit une matrice diagonale. En remplaçant A par $UD{}^tV$, on obtient $B = V(D^+ - (D^2 + \lambda^2 E)^{-1}D){}^tU$, d'où $\|B\| = \|D^+ - (D^2 + \lambda^2 E)^{-1}D\|$. La matrice $D^+ + (D^2 + \lambda^2 E)^{-1}D$ est diagonale et ses éléments non nuls sont $\alpha_i^{-1} - \alpha_i(\alpha_i^2 + \lambda^2)^{-1}$, $i = 1, \ldots, r$. Le nombre r est ici égal à Rg A, tandis que α_i sont les nombres singuliers non nuls de la matrice A. Selon le théorème 2 du § 1, ch. XI, les éléments diagonaux de la matrice $D^+ - (D^2 + \lambda^2 E)^{-1}D$ sont ses nombres singuliers. Le plus grand d'entre eux est égal à la norme spectrale de la matrice, de sorte que

$$\|A^+ - ({}^tAA + \lambda^2 E)^{-1}({}^tA)\| = \alpha_r^{-1} - \alpha_r(\alpha_r^2 + \lambda^2)^{-1} = \frac{\lambda^2}{\alpha_r(\alpha_r^2 + \lambda^2)}.$$

Donc,

$$\|x_0 - x_\lambda\| \leqslant \frac{\lambda^2\|b\|}{\alpha_r(\alpha_r^2 + \lambda^2)} \leqslant \frac{\lambda^2\|b\|}{\alpha_r^3} = \lambda^2 q, \tag{12}$$

où q est une constante.

Maintenant, en s'appuyant sur les majorations (11) et (12), on obtient

$$\|\tilde{x}_\lambda - x_0\| \leqslant \|\tilde{x}_\lambda - x_\lambda\| + \|x_\lambda - x_0\| \leqslant \frac{\delta}{\lambda^2} p(\delta, \lambda) + \lambda^2 q. \tag{13}$$

Admettons que la précision δ des données initiales soit fixée. Alors, lorsque λ décroît, le premier terme dans la majoration (13) augmente et le second terme diminue. Il est donc possible de choisir une valeur de λ minimisant $\|x_\lambda - x_0\|$.

La formule (12) permet de majorer la fonction $p(\delta, \lambda)$. En effet, il découle de (12) que

$$\|x_\lambda\| \leqslant \|x_0\| + \lambda^2 q.$$

En portant cette expression dans celle de $p(\delta, \lambda)$, on obtient

$$p(\delta, \lambda) \leqslant (c + \delta)(\|x_0\| + \lambda^2 q) + (\|b\| + \|A\| + \delta)\delta = \\ = c' + c''\delta + q(c + \delta)\lambda^2,$$

où c' et c'' sont des constantes strictement positives. Donc,

$$\frac{\delta}{\lambda^2} p(\delta, \lambda) \leqslant \frac{\delta}{\lambda^2}(c' + c''\delta) + \delta q(c + \delta).$$

Si on en tient compte dans la majoration (13), on trouve que

$$\|\tilde{x}_\lambda - x_0\| \leqslant \frac{\delta}{\lambda^2}(c' + c''\delta) + \delta q(c + \delta) + \lambda^2 q = \frac{s}{\lambda^2} + \lambda^2 q + t.$$

Cherchons la valeur de λ^2 pour laquelle le second membre de l'égalité devient minimal. En dérivant par rapport à λ^2, on obtient l'équation $-\frac{s}{\lambda^2} + q = 0$ qui donne la valeur cherchée

$$\lambda_1^2 = \sqrt{\frac{s}{q}} = \sqrt{\delta}\left(\frac{c' + c''\delta}{q}\right)^{1/2}. \tag{14}$$

On voit facilement que pour cette valeur de λ^2 la norme de la différence $\tilde{x}_\lambda - x_0$ est une grandeur d'ordre $\sqrt{\delta}$.

On ne discutera pas en détail la formule (14). Son importance n'est pas si grande vu que son second membre contient des inconnues ainsi que des quantités $\|x_0\|$, α_r et autres, difficilement calculables. Toutefois la formule (14) permet d'énoncer la proposition suivante.

Proposition 1. *On peut trouver la pseudo-solution du système* (1) *avec une erreur d'ordre $\sqrt{\delta}$, où δ définit l'erreur sur les données initiales suivant les formules* (9).

3. Calcul de la matrice pseudo-inverse. On étudiera maintenant les méthodes de calcul pratique de la matrice pseudo-inverse. Toutes les difficultés liées à la recherche de la matrice inverse d'une matrice régulière se conservent aussi dans ce calcul, mais il s'y ajoute d'autres difficultés. On a limité au chapitre XIII la classe des matrices en posant que les matrices étudiées ne sont pas quasi singulières. Ici on ne peut imposer à une matrice de telles restrictions, de sorte qu'il se pose le problème de définir le rang de la matrice. Il a été noté à la p. 382 que si on tient compte des erreurs d'arrondi et des erreurs sur les données initiales, le rang de la matrice ne peut être établi avec certitude pour toutes d'entre elles. Or le résultat des calculs dépend essentiellement de la valeur du rang de la matrice donnée. Considérons l'exemple suivant.

Soit

$$A = \begin{Vmatrix} 1 & 0 \\ 0 & 10^{-6} \end{Vmatrix}.$$

Alors

$$A^{-1} = \begin{Vmatrix} 1 & 0 \\ 0 & 10^{6} \end{Vmatrix}.$$

Si on néglige l'élément 10^{-6}, on obtient la matrice

$$B = \begin{Vmatrix} 1 & 0 \\ 0 & 0 \end{Vmatrix}$$

et sa pseudo-inverse

$$B^+ = \begin{Vmatrix} 1 & 0 \\ 0 & 0 \end{Vmatrix}$$

qui diffère sensiblement de A^{-1}.

L'accroissement de précision et une meilleure sélection d'algorithmes peuvent restreindre la classe des matrices de rang indéterminé, mais ce problème n'a pas de solution de principe.

Dans l'exposé qui suivra, on supposera, en s'appuyant sur certaines considérations liées peut-être à la position du problème où il a fallu calculer une matrice pseudo-inverse, qu'on est en mesure d'établir le rang de la matrice donnée.

Dans la plupart des cas, les méthodes de recherche de la matrice pseudo-inverse d'une matrice donnée A s'appuient sur la décomposition de A en produit de facteurs. Les propositions 8 et 10 du § 1 nous offrent deux cas, où de $A = BC$ on obtient $A^+ = C^+B^+$. Ces cas-là sont utilisés dans les calculs.

4. Obtention directe de la décomposition squelettique d'une matrice. Cette décomposition a été définie au point 3 du § 1. La proposition 10 du § 1 nous montre comment elle peut être utilisée à la pseudo-inversion de la matrice. Voyons comment cette décomposition peut être obtenue.

Soit une matrice A à m lignes et n colonnes. S'il est établi que cette matrice est de rang r, ses r colonnes peuvent être transformées par des opérations élémentaires sur les lignes en r premières colonnes de la matrice unité. Soient $j_1, \ldots, j_r$ les numéros des colonnes principales. On négligera, par hypothèse du rang de la matrice, les éléments des $m - r$ dernières lignes.

Désignons par C la matrice de type (r, n) composée des r premières lignes de la matrice transformée.

Soit B la matrice de type (m, r) composée des colonnes de la matrice A de numéros $j_1, \ldots, j_r$, autrement dit des colonnes principales. Il s'avère que

$$A = BC$$

et que cette décomposition est squelettique.

En effet, les colonnes principales (de numéros $j_1, \ldots, j_r$) de la matrice C sont les colonnes de la matrice unité d'ordre r. Aussi les éléments de la j-ième colonne de C sont-ils les coefficients dans la décomposition de cette colonne suivant les colonnes principales. Or on sait que les relations linéai-

res entre les colonnes de la matrice sont invariantes par les opérations élémentaires sur les lignes. Donc, la j-ième colonne de la matrice A se décompose suivant ses colonnes principales avec les mêmes coefficients. Il ne reste qu'à se rappeler que la j-ième colonne du produit BC est une combinaison linéaire des colonnes de B avec les coefficients égaux aux éléments de la j-ième colonne de C.

Ce procédé d'obtention de la décomposition squelettique présente plusieurs défauts évidents : il oblige de prendre une décision sur la petitesse des $m - r$ dernières lignes de la matrice transformée. De plus, le calcul de la matrice pseudo-inverse suivant la formule de la proposition 10, § 1, nécessite un grand nombre d'opérations arithmétiques et, à ce qu'il paraît, il n'y a aucune possibilité de se servir de la matrice de transformation obtenue avec la décomposition de la matrice A. La matrice B ne présente aucune structure spéciale outre l'indépendance linéaire de ses colonnes. Quant à la matrice C, on expliquera plus loin au point 7 comment on peut utiliser sa structure spéciale.

5. La *QR*-décomposition des matrices rectangulaires. Rappelons que pour trouver la QR-décomposition des matrices carrées par la méthode des symétries étudiée au § 3 du ch. XIII, on a successivement transformé les colonnes de numéros 1, 2, 3, etc. de la matrice en des colonnes correspondant à la forme triangulaire supérieure. Chaque transformation représentait la multiplication à gauche par une matrice de symétrie et ne modifiait pas la forme des colonnes déjà transformées. Si la i-ième colonne de la matrice était une combinaison linéaire des colonnes précédentes, ses éléments situés sur la diagonale et au-dessous d'elle s'annulaient après les transformations de ces colonnes. Une telle colonne ne nécessitait plus de transformation, et on omettait le facteur correspondant.

Le fait que la matrice qu'on transformait était carrée ne jouait au fond aucun rôle dans le processus. En transformant de la sorte une matrice A de type (m, n), on obtient la décomposition

$$A = QR,$$

où Q est une matrice orthogonale d'ordre m, et R une matrice de type (m, n) dont les éléments ρ_{ij} satisfont à la condition $\rho_{ij} = 0$ pour $i > j$. On dira que cette décomposition est la *Qr-décomposition*, en se servant de la lettre r pour rappeler que la matrice R n'est en général pas carrée.

On obtient une autre généralisation de la QR-décomposition si l'on se souvient de la QR-décomposition par la méthode d'orthogonalisation de Gram-Schmidt (voir p. 422). Soit A une matrice de type (m, n), avec $m > n$, et soit $\text{Rg}\, A = n$, ce qui veut dire que les colonnes de A sont linéairement indépendantes. En assimilant ces colonnes à n vecteurs de l'espace m-dimensionnel, on peut leur appliquer la méthode d'orthogonalisation et

construire une matrice triangulaire supérieure U, telle que les colonnes de la matrice $Q = AU$ soient orthogonales et normées par rapport au produit scalaire ordinaire dans l'espace des colonnes. Ainsi, la matrice Q de type (m, n) peut être considérée comme sous-matrice d'une matrice orthogonale d'ordre m. Désignons U^{-1} par R et l'on obtient la décomposition $A = QR$, où R est une matrice triangulaire supérieure carrée d'ordre n et Q la matrice décrite plus haut. Cette décomposition sera appelée *qR-décomposition*.

A la p. 422 on a noté que le processus d'orthogonalisation de Gram-Schmidt exposé sous sa forme habituelle n'est pas stable par rapport aux erreurs d'arrondi. Vu l'importance de la qR-décomposition pour la pseudo-inversion des matrices, et du processus d'orthogonalisation pour la qR-décomposition, on étudiera ici la méthode d'orthogonalisation modifiée ayant une plus grande stabilité.

6. Méthode de réorthogonalisation. Supposons qu'on doit orthogonaliser n matrices-colonnes à m éléments : $a_1, \ldots, a_n$ $(m > n)$. On sait que la méthode usuelle consiste à remplacer la première matrice-colonne par $q_1 = a_1/\|a_1\|$, où $\| * \|$ est la norme euclidienne. Ensuite, si $q_1, \ldots, q_k$ sont déjà construites, on pose

$$u = a_{k+1} - \sum_{i=1}^{k} c_i q_i,$$

où les coefficients c_i doivent vérifier les conditions ${}^t q_1 u = 0, \ldots, {}^t q_k u = 0$. Ces conditions peuvent être écrites sous forme d'égalités

$$\sum_{i=1}^{k} c_i ({}^t q_j) q_i = {}^t q_j a_{k+1} \tag{15}$$

pour tous les $j = 1, \ldots, k$. Notons Γ la matrice de ce système d'équations linéaires, c'est-à-dire la matrice d'éléments ${}^t q_j q_i$. Si les calculs sont faits exactement, les matrices-colonnes $q_1, \ldots, q_k$ sont orthonormées, Γ est une matrice unité et $c_i = {}^t q_i a_{k+1}$, d'où

$$u = a_{k+1} - \sum_{i=1}^{k} ({}^t q_i a_{k+1}) q_i. \tag{16}$$

Ensuite, on pose $q_{k+1} = u/\|u\|$ et on passe au calcul de q_{k+2}.

Cependant, si les calculs sont effectués de façon approchée, Γ n'est qu'une approximation de E et $c_i = {}^t q_i a_{k+1}$ n'est déjà plus solution du système (15). Ceci étant, le processus décrit plus haut aboutit au système des matrices-colonnes $q_1, \ldots, q_k$ qui, avec une grande précision, engendre le même sous-espace que $a_1, \ldots, a_n$. L'instabilité notée plus haut de la méthode d'orthogonalisation se traduit dans le fait que les matrices-

colonnes calculées q_i peuvent s'avérer de loin non orthogonales et l'altération due aux erreurs d'arrondi est d'autant plus grande que les matrices-colonnes a_i sont plus proches des colonnes linéairement dépendantes. Pour éviter cet inconvénient, on doit résoudre le système (15) sans exiger que les matrices-colonnes calculées q_i soient orthonormées. Dans ce cas, les erreurs introduites avec le calcul de $q_1, \ldots, q_k$ n'influent pas si fortement sur le calcul de q_{k+1}. On peut se servir de la méthode itérative simple (comp. § 4, ch. XIII), qui converge ici rapidement, vu que la matrice Γ est proche de la matrice unité.

On décrira ici le procédé de construction de la matrice-colonne q_{k+1}, appelé *méthode de réorthogonalisation*.

Soit u la matrice-colonne obtenue par la formule (16) et qu'on suppose non orthogonale à $q_1, \ldots, q_k$. On la note $u^{(1)}$ et l'on construit la matrice-colonne $u^{(2)}$ suivant la formule

$$u^{(2)} = u^{(1)} - \sum_{i=1}^{k} ({}^t q_i u^{(1)}) q_i.$$

D'une façon générale on peut construire la suite $u^{(0)} = a_{k+1}, u^{(1)}, u^{(2)}, \ldots$ en accord avec la relation de récurrence

$$u^{(s+1)} = u^{(s)} - \sum_{i=1}^{k} ({}^t q_i u^{(s)}) q_i. \tag{17}$$

Considérons pour un terme arbitraire de cette suite le produit ${}^t q_j u^{(s+1)}$ pour un $j \leqslant k$. On a

$${}^t q_j u^{(s+1)} = {}^t q_j u^{(s)} - \sum_{i=1}^{k} ({}^t q_i u^{(s)})({}^t q_j q_i).$$

Cela signifie que la matrice-ligne

$$p^{(s+1)} = \| {}^t q_1 u^{(s+1)}, \ldots, {}^t q_k u^{(s+1)} \|$$

vérifie la relation

$$p^{(s+1)} = p^{(s)}(E - \Gamma).$$

La suite de $p^{(s)}$ converge vers la matrice-ligne nulle et de plus très vite, car Γ est proche de E. En effet, il ressort des propriétés de la norme euclidienne que

$$\|p^{(s+1)}\| \leqslant \|p^{(s)}\| \cdot \|E - \Gamma\|_E \leqslant \|a_{k+1}\| \cdot \|E - \Gamma\|_E^{s+1}.$$

Démontrons maintenant que la suite de $u^{(s)}$ converge. En additionnant membre à membre les égalités de la forme (17) pour tous les $s = \alpha, \ldots, \beta$,

on obtient

$$u^{(\beta)} = u^{(\alpha)} - \sum_{i=1}^{k} ({}^t q_i u^{(\beta-1)} + \ldots + {}^t q_i u^{(\alpha)}) q_i, \tag{18}$$

d'où

$$\|u^{(\beta)} - u^{(\alpha)}\| \leqslant \sum_{i=1}^{k} |{}^t q_i u^{(\beta-1)} + \ldots + {}^t q_i u^{(\alpha)}| \cdot \|q_i\| \leqslant$$

$$\leqslant \sum_{i=1}^{k} \sum_{s=\alpha}^{\beta-1} |{}^t q_i u^{(s)}| \cdot \|q_i\|.$$

En posant $\|q_i\| \leqslant 2$ et en modifiant l'ordre de sommation, on obtient

$$\|u^{(\beta)} - u^{(\alpha)}\| \leqslant 2 \sum_{s=\alpha}^{\beta-1} \|p^{(s)}\|_1.$$

Vu que la l-norme $\| * \|_1$ ne dépasse pas la norme euclidienne, on a

$$\|u^{(\beta)} - u^{(\alpha)}\| \leqslant 2\|a_{k+1}\| \sum_{s=\alpha}^{\beta-1} \|E - \Gamma\|_E^s \leqslant \frac{2\|a_{k+1}\| \cdot \|E - \Gamma\|_E^{\alpha+1}}{1 - \|E - \Gamma\|_E}.$$

Il n'est maintenant pas difficile de montrer que pour $\varepsilon > 0$ donné, on peut trouver un numéro γ de manière que $\|u^{(\beta)} - u^{(\alpha)}\| < \varepsilon$, dès que $\alpha, \beta > \gamma$. Ainsi, on voit que la suite de $u^{(s)}$ satisfait au critère de Cauchy.

Ensuite, en posant $\alpha = 0$ dans la formule (18), on trouve que pour certains coefficients $c_i^{(\beta)}$

$$u^{(\beta)} = a_{k+1} - \sum_{i=1}^{k} c_i^{(\beta)} q_i.$$

Aussi la limite v de la suite $\{u^{(s)}\}$ vérifie-t-elle l'égalité

$$v = a_{k+1} - \sum_{i=1}^{k} c_i q_i,$$

où $c_i = \lim_{\beta \to \infty} c_i^{(\beta)}$. Ceci montre que $v \neq \mathbf{0}$, si a_{k+1} n'est pas une combinaison linéaire des matrices-colonnes $a_1, \ldots, a_k$.

On a vu que $p^{(s)} \to \mathbf{0}$ $(s \to \infty)$. Aussi la matrice-colonne v est-elle orthogonale à toutes les $q_1, \ldots, q_k$.

Il va de soi que pratiquement au lieu de v on prend $u^{(s)}$ de numéro s suffisamment grand. Vu que la norme $\|E - \Gamma\|_E$ est petite, la suite converge rapidement et on peut se limiter en général au calcul de $u^{(2)}$.

Après avoir construit avec une précision nécessaire la matrice-colonne v non nulle orthogonale aux matrices-colonnes $q_1, \ldots, q_k$, on pose $q_{k+1} = v/\|v\|$ et l'on passe au calcul de q_{k+2}.

On avait supposé que les matrices-colonnes $a_1, \ldots, a_n$ qu'on devait orthogonaliser étaient linéairement indépendantes. Laissons tomber cette hypothèse. Posons que la matrice-colonne a_{k+1} s'exprime linéairement au moyen de $a_1, \ldots, a_k$. Alors la matrice-colonne correspondante v devient nulle. Elle ne peut pas naturellement être normée et on ne la joint pas à l'ensemble des matrices-colonnes orthonormées construites. A l'étape suivante, on procède à l'orthogonalisation de la matrice-colonne a_{k+2} par rapport aux matrices-colonnes $q_1, \ldots, q_k$ déjà existantes.

Remarquons que si le rang de la matrice est inconnu, on doit chaque fois prendre une décision : peut-on admettre que la matrice-colonne v est nulle si ses éléments calculés sont de faible valeur.

Toutefois, si cette difficulté est surmontée d'une façon quelconque, on peut trouver pour la matrice A de type (m, n), avec $m > n$, par la méthode d'orthogonalisation, la qR-décomposition $A = QR$, où R est une matrice triangulaire supérieure régulière et Q une matrice dont les colonnes non nulles sont orthonormées. Si $m < n$, les raisonnements demeurent les mêmes, sauf que la matrice Q contient nécessairement quelques colonnes nulles.

7. Applications de la qR-décomposition. Montrons d'abord que la qR-décomposition d'une matrice quelconque A permet d'obtenir sa décomposition squelettique. Posons pour simplifier les notations que les r premières colonnes dans la matrice A sont principales. S'il n'en est pas ainsi, on peut toujours permuter les colonnes, ce qui n'entraînera selon la proposition 8 du § 1 qu'une permutation des lignes dans la matrice A^+. Ainsi donc, on pose que la matrice A est décomposée en blocs

$$A = \|B, S\|,$$

le bloc B étant de type (m, r) et de rang r.

En calculant la qR-décomposition de la matrice A par orthogonalisation, on aboutit à la matrice triangulaire régulière R d'ordre n, telle que $AR = Q$, où Q est de la forme

$$Q = \|Q_1, O\|,$$

le bloc Q_1 étant composé de r colonnes orthonormées.

Toute colonne de numéro j dans la matrice R est composée des coefficients avec lesquels la colonne de même numéro j dans la matrice Q se

décompose suivant les colonnes de la matrice A. Pour $j > r$, les colonnes de Q sont nulles. Aussi la j-ième colonne de la matrice R pour $j > r$ peut-elle être prise sous la forme

$$'\| \rho_{1j} \dots \rho_{rj}\ 0 \dots 1 \dots 0 \|,$$

où $\rho_{1j}, \dots, \rho_{rj}$ sont les opposés des coefficients qui figurent dans la décomposition de la j-ième colonne de A suivant les colonnes principales. Désignons par R'' la sous-matrice de la matrice R, formée à partir de ses $n - r$ dernières colonnes. Il ressort de ce qu'il vient d'être dit que R'' est de la forme

$$R'' = \left\| \begin{matrix} -U \\ E_{n-r} \end{matrix} \right\|,$$

où U est un bloc à r lignes et $n - r$ colonnes. Vu que $AR'' = O$, on a $B(-U) + S = O$, ou $S = BU$. Il s'ensuit que $A = \|B, BU\| = B\|E_r, U\|$, ou bien

$$A = BC,$$

avec $C = \|E_r, U\|$. Cette décomposition est évidemment squelettique.

Les matrices calculées R et Q permettent aussitôt d'obtenir la matrice pseudo-inverse pour le premier facteur de la décomposition squelettique. En effet, soit R_1' la sous-matrice de R, formée à partir de ses r premières lignes et colonnes. On vérifie aisément que

$$BR_1' = Q_1,$$

où Q_1 est une sous-matrice composée des colonnes non nulles de la matrice Q. Cette décomposition est aussi squelettique. Donc, $Q_1^+ = (R_1')^{-1}B^+$ et $B^+ = R_1'Q_1^+$. Or les colonnes de Q_1 sont linéairement indépendantes et ${}'Q_1Q_1 = E$. Donc, $Q_1^+ = ({}'Q_1Q_1)^{-1}({}'Q_1) = {}'Q_1$. Finalement,

$$B^+ = R_1'({}'Q_1).$$

Pour trouver C^+, il faut encore une fois recourir à la méthode d'orthogonalisation et rechercher une matrice triangulaire régulière Z d'ordre r telle que ${}'CZ = P$, où P est une matrice à colonnes orthonormées. Ceci est bien possible, vu que ies lignes de la matrice C sont linéairement indépendantes. En raisonnant alors comme dans le cas de B, on obtient

$$({}'C)^+ = Z\,{}'P$$

et

$$C^+ = P\,{}'Z.$$

Notons en conclusion qu'on a utilisé pour les matrices A et C le résultat d'orthogonalisation de leurs colonnes et non pas la qR-décomposition. Si

on connaît la qR-décomposition obtenue d'une autre façon quelconque, il ne reste plus que d'inverser la matrice triangulaire supérieure, ce qui ne présente pas de grandes difficultés.

8. Seconde forme de décomposition singulière. Le procédé le plus efficace, bien que relativement laborieux, de résolution des problèmes liés à la recherche de la pseudo-solution normale et à la pseudo-inversion des matrices, est l'obtention de la décomposition singulière. En maintes occasions, la recherche de la décomposition singulière peut remplacer la recherche de la matrice pseudo-inverse. La décomposition singulière introduite au § 1 du ch. XI peut revêtir une autre forme qu'on va étudier.

Désignons par E_n^r une matrice de type (r, n) composée des r premières lignes de la matrice unité d'ordre n. Quelle que soit la matrice V, le produit $E_n^r V$ (s'il est défini) est composé des r premières lignes de la matrice V. Si le produit SE_n^r est défini, il représente la matrice S, à r colonnes de laquelle viennent se joindre à droite $n - r$ colonnes nulles.

Considérons une matrice A de type (m, n) et de rang r. Sa décomposition singulière est de la forme $A = {}^tUDV$, où D est une matrice diagonale de type (m, n) dont la forme est

$$\begin{Vmatrix} \Lambda & O \\ O & O \end{Vmatrix},$$

avec Λ une matrice diagonale carrée d'ordre r. La matrice D peut être représentée sous forme de produit ${}^t(E_m^r)\Lambda E_n^r$. Alors,

$$A = {}^tU\,{}^t(E_m^r)\Lambda E_n^r V = {}^t(E_m^r U)\Lambda(E_n^r V).$$

Selon ce qui a été dit plus haut, $E_n^r V$ est une matrice composée des r premières lignes de V, et ${}^t(E_m^r U)$ est une matrice composée des r premières colonnes de la matrice ${}^tU$. Désignons ces matrices respectivement par V_1 et ${}^tU_1$. Les lignes de V_1 et U_1 (colonnes de ${}^tU_1$) sont orthonormées. On a obtenu la décomposition

$$A = {}^tU_1 \Lambda V_1, \tag{19}$$

qu'on appellera *seconde forme de décomposition singulière*.

La seconde forme de décomposition singulière contient des matrices de moindres dimensions que les matrices de la première forme : il n'est pas nécessaire de calculer et de retenir les composantes des vecteurs de bases singulières, qui correspondent aux nombres singuliers nuls.

Donnons une interprétation sommaire de la décomposition (19) en termes d'applications linéaires.

Soient $\mathscr{E}_n$ et $\bar{\mathscr{E}}_m$ des espaces euclidiens et $\mathbf{A}$ une application linéaire de $\mathscr{E}_n$ dans $\bar{\mathscr{E}}_m$. Définissons le projecteur orthogonal de $\mathscr{E}_n$ sur Im $\mathbf{A}^*$ comme une application $\mathbf{P} : \mathscr{E}_n \to$ Im $\mathbf{A}^*$ associant à chaque vecteur $x \in \mathscr{E}_n$ sa projection

orthogonale sur Im $\mathbf{A}^*$ (comp. point 4, § 3, ch. XII). Si $\mathscr{E}_n$ et Im $\mathbf{A}^*$ sont rapportés à des bases choisies, l'application $\mathbf{P}$ se définit par une matrice P de type (r, n).

Vu que $(\mathrm{Im}\, \mathbf{A}^*)^{\perp} = \mathrm{Ker}\, \mathbf{A}$, on a pour tout $x \in \mathscr{E}_n$

$$\mathbf{A}(x) = \mathbf{A} \circ \mathbf{P}(x).$$

On le vérifie facilement en écrivant x sous la forme $x' + x''$, où $x' \in \mathrm{Im}\, \mathbf{A}^*$ et $x'' \in \mathrm{Ker}\, \mathbf{A}$.

Considérons la restriction $\mathbf{A}_0$ de l'application $\mathbf{A}$ au sous-espace Im $\mathbf{A}^*$. Selon la proposition 1 du § 2, $\mathbf{A}_0$ est une bijection de Im $\mathbf{A}^*$ sur Im $\mathbf{A}$. Soit en outre $\mathbf{I}$ l'injection canonique de Im $\mathbf{A}$ dans $\bar{\mathscr{E}}_m$ (voir pour plus de détails sur cette application p. 352). On peut maintenant représenter l'application $\mathbf{A}$ sous forme de produit

$$\mathbf{A} = \mathbf{I} \circ \mathbf{A}_0 \circ \mathbf{P}. \qquad (20)$$

L'expression matricielle de ce produit devient possible après le choix de quatre bases dans les espaces $\mathscr{E}_n$, Im $\mathbf{A}^*$, Im $\mathbf{A}$ et $\bar{\mathscr{E}}_m$. Rapportons les espaces $\mathscr{E}_n$ et $\bar{\mathscr{E}}_m$ à deux bases orthonormées quelconques et notons A la matrice de l'application $\mathbf{A}$ par rapport à ces bases. Ensuite, rapportons les espaces Im $\mathbf{A}^*$ et Im $\mathbf{A}$ à deux bases formées respectivement des vecteurs de la première et de la seconde base singulière de $\mathbf{A}$, correspondant aux nombres singuliers non nuls (il y a exactement r vecteurs dans chaque base). La matrice Λ de l'application $\mathbf{A}_0$ par rapport à ces bases est une matrice diagonale avec nombres singuliers non nuls sur la diagonale.

Il est évident que dans les bases ainsi choisies la matrice de l'application $\mathbf{I}$ présente des colonnes orthonormées. En passant dans l'espace $\bar{\mathscr{E}}_m$ à la seconde base singulière, on peut constater que pour la base choisie dans Im $\mathbf{A}$ l'application $\mathbf{P}$ a pour matrice E_m^r. Donc, pour la base initiale dans $\bar{\mathscr{E}}_m$ la matrice de $\mathbf{P}$, égale à $E_m^r V$, possède des lignes orthonormées.

Ainsi, la seconde forme de décomposition singulière peut être obtenue comme une expression analytique de la décomposition (20).

Remarque. En rapport avec la décomposition (20), il faut noter que dans la définition de l'application pseudo-inverse, faite au début du § 2, il aurait fallu écrire devant $\mathbf{A}_0^{-1}$ l'injection canonique $\mathbf{I}$ de l'espace Im $\mathbf{A}^*$ dans $\mathscr{E}_n$. On s'abstient de le faire vu que dans ce cas, comme dans les autres, aucune difficulté ne se présente si cette injection est tout simplement sous-entendue. Mais il arrive des situations (dont l'une vient d'être rencontrée) quand l'introduction de ce facteur « auxiliaire » rend la situation plus claire.

La voie la plus naturelle de construction de la décomposition singulière d'une matrice A est de résoudre complètement le problème des valeurs propres pour la matrice ${}^tAA$. Mais pratiquement cette voie est peu acceptable.

En effet, la recherche de la décomposition singulière pose l'un des plus importants problèmes d'existence du nombre singulier nul et de sa multiplicité. Si les nombres singuliers sont obtenus comme racines carrées des nombres caractéristiques de ${}^tAA$, la résolution du problème se complique car les carrés des nombres singuliers sont plus difficiles à distinguer de zéro que les nombres eux-mêmes. De plus, le calcul de la racine carrée des nombres proches de zéro est entaché de l'erreur relative très grande.

Si la matrice ${}^tAA$ est tout de même utilisée, il découle de ce qu'on vient de dire que les éléments de cette matrice doivent être calculés avec une très grande précision.

Aussi en pratique (voir Forsythe, Malcolm et Moler [11], Wilkinson et Reinsch [43]) utilise-t-on la méthode qu'on décrira pour le cas de $m \geqslant n$. (Pour $m < n$, la même méthode peut être appliquée à la matrice transposée.)

D'abord, par la méthode des symétries, on construit des matrices orthogonales U et V telles que la matrice $A^{(0)} = {}^tUAV$ soit bidiagonale, c'est-à-dire que ses éléments $a_{ij}^{(0)}$ peuvent ne pas être nuls si $i = j$ ou $i + 1 = j$. La réduction à la forme bidiagonale ne diffère en rien de celle qui a été décrite pour les matrices carrées au § 5 du ch. XIII.

Ensuite, on transforme la matrice $A^{(0)}$ par des matrices orthogonales $S^{(k)}, P^{(k)}$, $k = 0, 1, 2, \ldots$, suivant les formules

$$A^{(k+1)} = S^{(k)} A^{(k)} P^{(k)}.$$

Les matrices $S^{(k)}$, $P^{(k)}$ sont choisies de manière que la suite de $A^{(k)}$ converge vers la matrice diagonale D. Ceci étant, chaque matrice $A^{(k)}$ est bidiagonale et $a_{i,i+1}^{(k)} \to 0$ pour $k \to \infty$.

Si $\bar{S} = \lim_{k\to\infty} S^{(k)} \ldots S^{(0)}$ et $\bar{P} = \lim_{k\to\infty} P^{(k)} \ldots P^{(0)}$, on a $D = \bar{S}UAV\bar{P}$ et la décomposition $A = {}^t(\bar{S}U)D{}^t(V\bar{P})$ est une décomposition singulière de A.

Etudions maintenant les matrices $S^{(k)}$ et $P^{(k)}$. Chacune d'elles est le produit de matrices de rotation d'un plan :

$$S^{(k)} = S_2 \ldots S_n$$

et

$$P^{(k)} = P_2, \ldots P_n,$$

S_i et P_i étant des rotations du plan engendré par e_{i-1} et e_i. Ces rotations sont choisies de manière que la matrice $B^{(k)} = {}^tA^{(k)}A^{(k)}$ se transforme en matrice $B^{(k+1)} = {}^tP^{(k)}B^{(k)}P^{(k)}$ qu'on obtient de $B^{(k)}$ en une étape du QR-algorithme avec décalage. Cependant, les matrices S_i et P_i ne peuvent être trouvées que d'après les éléments de la matrice $A^{(k)}$, quant à la matrice $B^{(k)}$, elle n'est ni utilisée ni calculée.

Tout comme le QR-algorithme, ce processus itératif converge rapidement et fournit un résultat exact.

9. Utilisation de la décomposition singulière. En groupant les facteurs dans la seconde forme de décomposition singulière :

$$A = {}^t U_1 (\Lambda V_1),$$

on obtient la décomposition squelettique de la matrice A. En effet, les matrices ${}^t U_1$ et ΛV_1 sont respectivement de type (m, r) et (r, n), de sorte qu'on peut écrire $A^+ = (\Lambda V_1)^+ ({}^t U_1)^+$, d'où $A^+ = (\Lambda V_1)^+ U_1$ car les colonnes de la matrice ${}^t U_1$ sont orthonormées.

Soit $B = \Lambda V_1$. Cette décomposition de la matrice de type (r, n) en produit des matrices de type (r, r) et (r, n) et de rang r est squelettique. Par conséquent, $B^+ = V_1^+ \Lambda^{-1} = {}^t V_1 \Lambda^{-1}$ et, en définitive, on a

$$A^+ = {}^t V_1 \Lambda^{-1} U_1.$$

Si l'on a obtenu la décomposition singulière de la matrice A, cette formule permet de passer à la pseudo-inverse de A presque sans calculs supplémentaires. Il va de soi que la décomposition singulière est plus difficile à obtenir que par exemple la qR-décomposition, mais du point de vue des calculs, elle présente d'importants avantages. A savoir, on a déjà souligné plusieurs fois que le problème le plus difficile dans la pseudo-inversion des matrices est celui de la détermination du rang de la matrice. Ce problème n'apparaît pas si on utilise la décomposition singulière.

Il se pose un autre problème beaucoup plus simple, quand il faut décider si un nombre donné est suffisamment petit pour qu'on puisse le négliger. Plus précisément, lesquels des nombres singuliers calculés sont en réalité nuls.

Pour chaque problème concret, en tenant compte de l'estimation de l'erreur sur les données initiales ainsi que de celle des erreurs d'arrondi introduites par le calcul des nombres singuliers, on fixe certain nombre ε en guise de frontière : tout nombre singulier inférieur à ε est considéré comme nul.

Ainsi, les dimensions des matrices dans la décomposition calculée (19) dépendent de la valeur de ε. Cette décomposition est la décomposition exacte d'une autre matrice A_ε. Le rang de A_ε dépend de ε. Ceci étant, $\mathrm{Rg}\, A_\varepsilon \leqslant \mathrm{Rg}\, A$ et le nombre conditionnel spectral vérifie l'inégalité

$$c(A_\varepsilon) \leqslant \frac{\alpha_1}{\varepsilon} \leqslant c(A),$$

où α_1 est le plus grand nombre singulier de la matrice A.

On voit que la matrice A_ε est en général mieux conditionnée que A, mais une partie de l'information contenue dans A peut se perdre. Si l'on

admet comme nul un nombre singulier dont la valeur exacte est strictement positive, on admet par là même qu'une ligne de A est linéairement dépendante des autres. Donc, en rendant ε plus grand que le nombre singulier minimal de A, on choisit dans A l'information la plus certaine dont la quantité est naturellement moins grande. Simultanément, en augmentant ε, on rend les notations plus simples et, partant, on économise la mémoire.

On peut utiliser ce résultat pour dégager la plus essentielle partie des données contenues dans la matrice. On rend pour cela ε suffisamment grand pour obtenir la matrice A_ε de rang admissible, à savoir : suffisamment petit pour que les données puissent être embrassées et suffisamment grand pour qu'elles présentent un intérêt.

10. Méthode de Gréville. Une toute autre approche à la recherche de la matrice pseudo-inverse s'appuie sur la construction suivante.

Supposons que la matrice A_1 est déduite de la matrice A par l'adjonction à droite la matrice-colonne $\boldsymbol{\alpha}$:

$$A_1 = \| A, \boldsymbol{\alpha} \| = \left\| \begin{matrix} a_{11} & \dots & a_{1n} & \alpha_1 \\ \dots & \dots & \dots & \dots \\ a_{m1} & \dots & a_{mn} & \alpha_m \end{matrix} \right\|.$$

Admettons d'abord que $AA^+\boldsymbol{\alpha} \neq \boldsymbol{\alpha}$ et considérons la matrice-colonne

$$\boldsymbol{\beta} = \frac{(E - AA^+)\boldsymbol{\alpha}}{\|(E - AA^+)\boldsymbol{\alpha}\|^2}, \tag{21}$$

où $\| * \|$ est la norme euclidienne.

PROPOSITION 2. *Si $AA^+\boldsymbol{\alpha} \neq \boldsymbol{\alpha}$, la matrice A_1^+ se déduit de la matrice $A^+(E - \boldsymbol{\alpha}\,'\boldsymbol{\beta})$ par l'adjonction en bas de la ligne $'\boldsymbol{\beta}$, c'est-à-dire est de la forme*

$$\left\| \begin{matrix} A^+(E - \boldsymbol{\alpha}\,'\boldsymbol{\beta}) \\ '\boldsymbol{\beta} \end{matrix} \right\|. \tag{22}$$

On le démontre le plus facilement en vérifiant les conditions de la proposition 12, § 1 pour la matrice (22). En la notant X, on obtient

$$A_1 X = \| A, \boldsymbol{\alpha} \| \left\| \begin{matrix} A^+(E - \boldsymbol{\alpha}\,'\boldsymbol{\beta}) \\ '\boldsymbol{\beta} \end{matrix} \right\| =$$

$$= AA^+(E - \boldsymbol{\alpha}\,'\boldsymbol{\beta}) + \boldsymbol{\alpha}\,'\boldsymbol{\beta} = AA^+ + (E - AA^+)\boldsymbol{\alpha}\,'\boldsymbol{\beta}.$$

Pour abréger les notations, désignons le dénominateur de l'expression (21) par λ. Alors,

$$A_1 X = AA^+ + (E - AA^+)\boldsymbol{\alpha}\,'\boldsymbol{\alpha}\,'(E - AA^+)\lambda^{-1},$$

d'où l'on voit que la matrice $A_1 X$ est symétrique.

Considérons maintenant le produit $A_1 X A_1$. On a

$$A_1 X A_1 = A_1 X \| A, \boldsymbol{\alpha} \| = \| A_1 X A, A_1 X \boldsymbol{\alpha} \|.$$

Calculons la matrice $A_1 X A$. Elle est égale à

$$A A^+ A + \lambda^{-1} (E - A A^+) \boldsymbol{\alpha}\, {}'\boldsymbol{\alpha} (E - A A^+) A,$$

car $E - AA^+$ est symétrique, tout comme AA^+.

Le premier terme vaut A. Chassons les parenthèses dans le second terme :

$$(E - AA^+) \boldsymbol{\alpha}\, {}'\boldsymbol{\alpha} (E - AA^+) A =$$
$$= \boldsymbol{\alpha}\, {}'\boldsymbol{\alpha} A - AA^+ \boldsymbol{\alpha}\, {}'\boldsymbol{\alpha} A - \boldsymbol{\alpha}\, {}'\boldsymbol{\alpha} AA^+ A + AA^+ \boldsymbol{\alpha}\, {}'\boldsymbol{\alpha} AA^+ A$$

et l'on voit qu'il est nul en vertu de l'identité $AA^+A = A$.

Calculons maintenant la matrice-colonne $A_1 X \boldsymbol{\alpha}$. Elle vaut

$$AA^+ \boldsymbol{\alpha} + \lambda^{-1} (E - AA^+) \boldsymbol{\alpha}\, {}'\boldsymbol{\alpha} (E - AA^+) \boldsymbol{\alpha}. \tag{23}$$

Pour simplifier cette expression, cherchons λ :

$$\lambda = {}'\boldsymbol{\alpha} (E - AA^+)(E - AA^+) \boldsymbol{\alpha} = {}'\boldsymbol{\alpha} (E - 2AA^+ + (AA^+)^2 \boldsymbol{\alpha} =$$
$$= {}'\boldsymbol{\alpha} (E - AA^+) \boldsymbol{\alpha}.$$

En simplifiant l'expression (23), on obtient

$$A_1 X \boldsymbol{\alpha} = AA^+ \boldsymbol{\alpha} + (E - AA^+) \boldsymbol{\alpha} = \boldsymbol{\alpha}.$$

Donc, $A_1 X A_1 = \| A, \boldsymbol{\alpha} \| = A_1$, ce qu'il fallait démontrer.

Ensuite, on peut vérifier que la matrice XA_1 se décompose en blocs suivants (qui sont des matrices de type (n, n), $(n, 1)$, $(1, n)$ et $(1, 1)$) :

$$XA_1 = \left\| \begin{matrix} A^+ (E - \boldsymbol{\alpha}\, {}'\boldsymbol{\beta}) A & A^+ (E - \boldsymbol{\alpha}\, {}'\boldsymbol{\beta}) \boldsymbol{\alpha} \\ {}'\boldsymbol{\beta} A & {}'\boldsymbol{\beta} \boldsymbol{\alpha} \end{matrix} \right\|. \tag{24}$$

Un calcul analogue au précédent montre que

$$XA_1 = \left\| \begin{matrix} A^+ A & 0 \\ 0 & 1 \end{matrix} \right\|,$$

et l'on voit que cette matrice est symétrique. En outre, en multipliant cette matrice par X, on obtient immédiatement $XA_1 X = X$. La proposition est démontrée.

Voyons maintenant le cas où $AA^+ \boldsymbol{\alpha} = \boldsymbol{\alpha}$. Posons

$$\boldsymbol{\gamma} = \frac{{}'(A^+) A^+ \boldsymbol{\alpha}}{1 + \| A^+ \boldsymbol{\alpha} \|^2} \tag{25}$$

et démontons la

PROPOSITION 3. *Si* $AA^+\boldsymbol{\alpha} = \boldsymbol{\alpha}$, *la matrice* A_1^+ *se déduit de la matrice* $A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma})$ *par l'adjonction en bas de la ligne* ${}^t\boldsymbol{\gamma}$, *c'est-à-dire vaut*

$$\left\| \begin{matrix} A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) \\ {}^t\boldsymbol{\gamma} \end{matrix} \right\|. \tag{26}$$

Par analogie à la proposition précédente, la démonstration consiste à vérifier les conditions de la proposition 12 du § 1. Notons X la matrice (26) et calculons A_1X :

$$A_1X = \|A, \boldsymbol{\alpha}\| \left\| \begin{matrix} A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) \\ {}^t\boldsymbol{\gamma} \end{matrix} \right\| = AA^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma} = AA^+$$

en vertu de la condition $AA^+\boldsymbol{\alpha} = \boldsymbol{\alpha}$. Ainsi, la matrice A_1X est symétrique. De façon analogue, indépendamment de la forme de $\boldsymbol{\gamma}$, on a

$$A_1XA_1 = AA^+A_1 = \|AA^+A, AA^+\boldsymbol{\alpha}\| = \|A, \boldsymbol{\alpha}\| = A_1.$$

Calculons la matrice XA_1. Elle est de la forme (24), à la seule différence qu'au lieu de $\boldsymbol{\beta}$ on a $\boldsymbol{\gamma}$. Transformons les expressions des blocs de cette matrice :

$$A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma})A = A^+A - A^+\boldsymbol{\alpha}\,{}^t\boldsymbol{\alpha}\,{}^t(A^+)A^+A\lambda^{-1},$$

où

$$\lambda = 1 + \|A^+\boldsymbol{\alpha}\|^2.$$

Or ${}^t(A^+)A^+A = {}^t(A^+)$ en vertu de l'identité (8) du § 1. Donc,

$$A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma})A = A^+A - (A^+\boldsymbol{\alpha})({}^t\boldsymbol{\alpha}\,{}^t(A^+))\lambda^{-1} =$$
$$= A^+A - \lambda^{-1}(A^+\boldsymbol{\alpha})\,{}^t(A^+\boldsymbol{\alpha}).$$

Il en découle que le bloc supérieur gauche est symétrique. Ensuite, en vertu de la même identité,

$${}^t\boldsymbol{\gamma}A = {}^t\boldsymbol{\alpha}\,{}^t(A^+)A^+A\lambda^{-1} = \lambda^{-1}({}^t\boldsymbol{\alpha})\,{}^t(A^+).$$

En outre,

$$A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma})\boldsymbol{\alpha} = A^+\boldsymbol{\alpha} - \lambda^{-1}A^+\boldsymbol{\alpha}\,{}^t\boldsymbol{\alpha}\,{}^t(A^+)A^+\boldsymbol{\alpha} =$$
$$= \lambda^{-1}(A^+\boldsymbol{\alpha}\lambda - A^+\boldsymbol{\alpha}\,{}^t\boldsymbol{\alpha}\,{}^t(A^+)A^+\boldsymbol{\alpha}) = \lambda^{-1}[A^+\boldsymbol{\alpha}(1 + {}^t\boldsymbol{\alpha}\,{}^t(A^+)A^+\boldsymbol{\alpha}) -$$
$$- A^+\boldsymbol{\alpha}\,{}^t\boldsymbol{\alpha}\,{}^t(A^+)A^+\boldsymbol{\alpha}] = \lambda^{-1}A^+\boldsymbol{\alpha}.$$

Cela montre que le bloc-ligne s'obtient par transposition du bloc-colonne.

On voit maintenant que la matrice XA_1 est symétrique. Etudions le produit XA_1X. Si Z_i, avec $i = 1, 2, 3, 4$, désignent les blocs de la matrice XA_1,

on a

$$XA_1X = \begin{Vmatrix} Z_1 & Z_2 \\ Z_3 & Z_4 \end{Vmatrix} \cdot \begin{Vmatrix} A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) \\ {}^t\boldsymbol{\gamma} \end{Vmatrix} \cdot \begin{Vmatrix} Z_1A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + Z_2({}^t\boldsymbol{\gamma}) \\ Z_3A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + Z_4({}^t\boldsymbol{\gamma}) \end{Vmatrix}.$$

En portant, au lieu de Z_i, leurs expressions en fonction de A, $\boldsymbol{\alpha}$ et $\boldsymbol{\gamma}$, on obtient

$$Z_1A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + Z_2({}^t\boldsymbol{\gamma}) = \\ = A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma})AA^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma})\boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}.$$

Chassons les parenthèses et tenons compte de ce que $AA^+\boldsymbol{\alpha} = \boldsymbol{\alpha}$ et que

$${}^t\boldsymbol{\gamma}\,AA^+ = \boldsymbol{\gamma}^{-1}({}^t\boldsymbol{\alpha})\,{}^t(A^+)A^+AA^+ = \lambda^{-1}({}^t\boldsymbol{\alpha})\,{}^tA^+)A^+ = {}^t\boldsymbol{\gamma}.$$

Alors l'expression à transformer devient

$$A^+ - A^+\boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma} = A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}).$$

Ensuite,

$$Z_3A^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + Z_4({}^t\boldsymbol{\gamma}) = \\ = {}^t\boldsymbol{\gamma}\,AA^+(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + {}^t\boldsymbol{\gamma}\,\boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma} = {}^t\boldsymbol{\gamma}(E - \boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma}) + {}^t\boldsymbol{\gamma}\,\boldsymbol{\alpha}\,{}^t\boldsymbol{\gamma} = {}^t\boldsymbol{\gamma}.$$

On a ainsi vérifié que $XA_1X = X$ et achevé la démonstration.

On peut utiliser les propositions 2 et 3 pour construire la matrice pseudo-inverse de la matrice donnée A. On considère pour cela n matrices $A_1, \ldots, A_n$, où la matrice A_t ($t = 1, \ldots, n$) est composée de t premières colonnes de la matrice A. Vu que la matrice A_1 comprend une seule colonne, A_1^+ s'obtient sans difficulté. Ensuite, on calcule successivement $A_2^+, A_3^+, \ldots$, jusqu'à ce qu'on n'obtienne $A_n^+ = A^+$.

Ce procédé est particulièrement commode si les données se complètent dans le temps et que l'on doive répéter pour la matrice A_{t+1} les calculs accomplis auparavant pour la matrice A_t si A_{t+1} se deduit de A_t par l'adjonction d'une colonne. Si l'on joint des lignes, la même méthode est utilisée pour la matrice transposée.

Il faut noter qu'en utilisant la méthode étudiée, on est toujours obligé de déterminer le rang de la matrice si on procède à la pseudo-inversion. Les difficultés qui y sont liées apparaissent juste au moment où l'on doit établir si l'égalité $AA^+\boldsymbol{\alpha} = \boldsymbol{\alpha}$ est vérifiée ou non et, en conséquence, choisir la formule nécessaire.

Il existe d'autres résultats ayant rapport aux liens entre la pseudo-inversion et la décomposition en blocs de la matrice, mais ils sont encore plus rebarbatifs (voir Albert [1]).

§ 4. Méthode des moindres carrés

1. Problème d'approximation de la fonction. Ce paragraphe est consacré aux applications des résultats obtenus auparavant sur les pseudo-solutions et les matrices pseudo-inverses.

Les systèmes contenant un grand nombre d'équations linéaires et un relativement faible nombre d'inconnues apparaissent dans le problème suivant d'une grande importance pratique dont un cas particulier a été envisagé dans l'introduction au § 1.

Supposons qu'on effectue une suite d'épreuves pour étudier la dépendance entre deux variables ξ et η, et que les résultats se présentent sous forme de m couples de valeurs

$$(\xi^1, \eta^1), \ldots, (\xi^m, \eta^m). \tag{1}$$

Il faut trouver une fonction $\eta = f(\xi)$ qui représenterait le mieux la dépendance réelle entre les variables. La fonction f est recherchée sous la forme d'une combinaison linéaire de fonctions données *a priori*

$$\varphi_1(\xi), \ldots, \varphi_n(\xi), \tag{2}$$

appelées *fonctions de base*.

On choisit ces fonctions suivant la nature du problème étudié. C'est ainsi que si l'on est fondé à supposer que la dépendance étudiée est périodique avec une période connue, on peut choisir pour fonctions de base les fonctions trigonométriques. Dans ce cas, la combinaison linéaire recherchée apparaît sous forme de somme partielle de la série de Fourier. On utilise souvent comme fonctions de base les fonctions puissances $(\xi)^k$ ($k = 0, \ldots, n-1$) et d'autres polynômes. Avec un tel choix, la fonction f sera un polynôme.

Les données expérimentales (1) contiennent des erreurs de mesure et, suivant la position du problème, d'autres erreurs aléatoires possibles. Aussi n'est-il pas exigé que toutes les égalités

$$\eta^i = f(\xi^i), \quad i = 1, \ldots, m, \tag{3}$$

soient vérifiées, c'est-à-dire que la courbe représentative de la fonction f passe par tous les points (1) du plan de coordonnées. Il n'est exigé que la courbe passe, en un certain sens, le plus près possible de tous ces points. On exige en général que la somme des carrés

$$(\eta^1 - f(\xi^1))^2 + \ldots + (\eta^m - f(\xi^m))^2 \tag{4}$$

soit minimale. Si la fonction f est choisie de cette manière, on dit qu'elle est choisie d'après la *méthode des moindres carrés*.

Soulignons la différence entre la position considérée du problème et le problème d'interpolation. Dans ce dernier, on recherche une fonction de

forme déterminée qui prend des valeurs données aux points indiqués. Dans le cas étudié, on suppose que les couples de nombres ne correspondent qu'approximativement à la dépendance réelle et on recherche une fonction de forme donnée qui correspond le mieux aux données expérimentales. Ceci étant, on ne s'attend pas à ce que la fonction obtenue représente exactement la dépendance réelle entre les variables, elle ne doit qu'approximer cette dépendance. En conséquence, dans un problème d'interpolation bien posé, le nombre de points est exactement égal au nombre de fonctions de base, tandis que dans la méthode des moindres carrés le nombre de points est en général beaucoup plus grand.

Si f est une combinaison linéaire des fonctions de base

$$f(\xi) = \theta^1 \varphi_1(\xi) + \ldots + \theta^n \varphi_n(\xi),$$

les égalités (3) représentent un système d'équations linéaires en $\theta^1, \ldots, \theta^n$:

$$\begin{gathered} a_{11}\theta^1 + \ldots + a_{1n}\theta^n = \eta^1, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}\theta^1 + \ldots + a_{mn}\theta^n = \eta^m, \end{gathered} \tag{5}$$

dont les coefficients sont les valeurs des fonctions de base aux points $\xi^1, \ldots$ $\ldots, \xi^m$, c'est-à-dire $a_{ij} = \varphi_j(\xi^i)$. Désignons la matrice du système par la lettre A.

L'écart défini pour un ensemble donné de coefficients $\theta^1, \ldots, \theta^n$ substitués dans le système possède les composantes

$$\eta^1 - f(\xi^1), \ldots, \eta^m - f(\xi^m),$$

le carré de sa norme ayant la forme (4). Par conséquent, tout ensemble de coefficients minimisant la somme des carrées (4) vérifie le système normal associé au système (5).

Il est toutefois important de souligner que, par rapport à la dépendance réelle entre les variables qu'on étudie, la solution du système normal n'est qu'une approximation des vraies valeurs des paramètres, ou comme on dit, n'est que leur *estimation*.

Si les colonnes de la matrice A du système (5) sont linéairement indépendantes, le système normal possède une solution unique qui constitue la pseudo-solution du système (5). En conséquence, le problème posé a également une solution unique.

Voyons dans quelle mesure on peut s'attendre à l'indépendance linéaire des colonnes de la matrice. Puisqu'on choisit les fonctions de base de la façon qu'il nous faut, il est donc naturel d'exiger qu'elles soient linéairement indépendantes. Rappelons que les fonctions $\varphi_1(\xi), \ldots, \varphi_n(\xi)$ sont dites ***linéairement dépendantes*** sur l'intervalle donné s'il existe une combinaison linéaire non triviale à coefficients constants de ces fonctions, qui est

identiquement égale à zéro sur cet intervalle. Si les fonctions de base sont linéairement dépendantes sur l'intervalle comprenant tous les points $\xi^1, \ldots$ $\ldots, \xi^m$, il en résulte évidemment l'indépendance linéaire des colonnes de la matrice. Dans ce cas, le choix de la pseudo-solution normale du système (5) pour coefficients $\theta^1, \ldots, \theta^n$ ne sera qu'un des possibles.

On voit facilement que l'indépendance linéaire des fonctions de base ne garantit pas l'indépendance linéaire des colonnes de la matrice. Par exemple, la matrice composée des valeurs des fonctions $\varphi_1(\xi) = 1$ et $\varphi_2(\xi) = (\xi)^2$ aux points $\xi^1 = 1$ et $\xi^2 = -1$ est

$$\left\| \begin{matrix} 1 & 1 \\ 1 & 1 \end{matrix} \right\|.$$

Mais si les fonctions de base sont linéairement indépendantes, la stricte dépendance linéaire entre les colonnes est un phénomène exceptionnel dans les conditions de la représentation approchée des nombres. Les colonnes de la matrice peuvent devenir voisines des colonnes linéairement dépendantes et la matrice A proche de la matrice de rang inférieur. Dans ce cas, comme on l'a vu, la pseudo-solution normale varie fortement pour de faibles modifications des données initiales.

C'est pourquoi, dans tous les cas, les calculs doivent être menés de manière à dévoiler l'instabilité numérique ou la non-unicité de la solution quand ces dernières ont lieu. Cela peut être fait de façon commode si en recherchant la pseudo-solution on se sert de la décomposition singulière de la matrice du système. On peut, comme il a été décrit au point 9 du § 3, introduire une frontière ε dépendant de la précision des calculs effectués et négliger les nombres singuliers de la matrice qui sont inférieurs à ε.

Estimons l'accroissement de la somme des carrés (4) engendré par l'élimination des nombres singuliers inférieurs à ε. Il est d'abord évident que toute solution $\beta^1, \ldots, \beta^n$ du système homogène normal ${}^tA A\boldsymbol{\beta} = \mathbf{0}$ peut être ajoutée aux coefficients $\theta^1, \ldots, \theta^n$ sans que la somme (4) soit modifiée (voir proposition 2, § 1).

Les solutions d'un système homogène normal constituent dans l'espace des matrices-colonnes à n éléments un sous-espace qu'on a désigné par la lettre $\mathscr{K}$ (il coïncide avec l'ensemble des solutions du système $A\boldsymbol{\beta} = \mathbf{0}$). En remplaçant par des zéros les nombres singuliers inférieurs à ε, on joint en fait à ce sous-espace les matrices-colonnes $\boldsymbol{\beta}_i$ pour lesquelles

$${}^tA\, A\boldsymbol{\beta}_i = \alpha_i^2 \boldsymbol{\beta}_i, \quad \alpha_i < \varepsilon, \tag{6}$$

et on peut considérer que $i = k, \ldots, r$.

Proposition 1. *Si à la solution* $\boldsymbol{\theta}$ *du système normal on ajoute une combinaison linéaire quelconque* $\boldsymbol{\gamma} = \gamma_k \boldsymbol{\beta}_k + \ldots + \gamma_r \boldsymbol{\beta}_r$ *des matrices-colonnes*

$\boldsymbol{\beta}_i$ *satisfaisant aux conditions* (6), *la somme des carrés* (4) *correspondant à la solution* $\boldsymbol{\theta}$ *s'accroît au plus de* $\varepsilon^2\|\boldsymbol{\gamma}\|^2$.

DÉMONSTRATION. La somme des carrés qu'on doit estimer est le carré de la norme de l'écart défini par la matrice-colonne $\boldsymbol{\theta} + \boldsymbol{\gamma}$, c'est-à-dire

$$\|\boldsymbol{\eta} - A(\boldsymbol{\theta} + \boldsymbol{\gamma})\|^2 = {}'[(\boldsymbol{\eta} - A\boldsymbol{\theta}) - A\boldsymbol{\gamma}][(\boldsymbol{\eta} - A\boldsymbol{\theta}) - A\boldsymbol{\gamma}].$$

En chassant les parenthèses, on a

$$\|\boldsymbol{\eta} - A(\boldsymbol{\theta} + \boldsymbol{\gamma})\|^2 = \|\boldsymbol{\eta} - A\boldsymbol{\theta}\|^2 + 2({}'\boldsymbol{\gamma})({}'AA\boldsymbol{\theta} - {}'A\boldsymbol{\eta}) + \|A\boldsymbol{\gamma}\|^2.$$

En tenant compte de ce que $\boldsymbol{\theta}$ est la solution du système normal, on obtient que la somme des carrés s'accroît de $\|A\boldsymbol{\gamma}\|^2$. Majorons cette grandeur.

Selon les relations (6), on a

$$\|A\boldsymbol{\gamma}\|^2 = {}'\boldsymbol{\gamma}\,{}'AA\boldsymbol{\gamma} = {}'\boldsymbol{\gamma}(\gamma_k({}'A)A\boldsymbol{\beta}_k + \ldots + \gamma_r({}'A)A\boldsymbol{\beta}_r) =$$
$$= {}'\boldsymbol{\gamma}(\gamma_k\alpha_k^2\boldsymbol{\beta}_k + \ldots + \gamma_r\alpha_r^2\boldsymbol{\beta}_r) \leqslant \varepsilon^2({}'\boldsymbol{\gamma})(\gamma_k\boldsymbol{\beta}_k + \ldots + \gamma_r\boldsymbol{\beta}_r) = \varepsilon^2\|\boldsymbol{\gamma}\|^2.$$

La proposition est démontrée.

La possibilité, fournie par la proposition 1, d'estimer simplement l'accroissement de la somme des carrés est un avantage important de la méthode permettant de trouver la pseudo-solution à l'aide de la décomposition singulière.

2. Régression linéaire. Le problème envisagé plus haut de l'estimation des paramètres de la fonction f d'après ses valeurs approchées est un cas particulier du problème plus général étudié en statistique mathématique. Soit y une variable aléatoire*). Supposons que son espérance mathématique $\mathbf{E}(y)$ dépend linéairement de n variables $a_1, \ldots, a_n$:

$$\mathbf{E}(y) = \theta^1 a_1 + \ldots + \theta^n a_n, \tag{7}$$

où les coefficients θ^i doivent être estimés sur le vu des valeurs observées de y.

Les variables a_j prennent m ensembles de valeurs qui forment une matrice A à m lignes et n colonnes :

$$\begin{array}{c} a_1^1, \ldots, a_n^1, \\ \ldots\ldots\ldots\ldots \\ a_1^m, \ldots, a_n^m. \end{array} \tag{8}$$

Les valeurs a_j^i sont considérées dans ce cas comme parfaitement connues.

On dit que la matrice A est une *matrice de régression* et que les variables a_j des *régresseurs*. Les régresseurs peuvent être des fonctions d'une seule ou

*) Avant d'aborder ce point, il est nécessaire de connaître les éléments de théorie des probabilités, par exemple, d'après le livre de Rozanov [33].

de plusieurs variables ; on peut aussi considérer que tous les a_j sont des variables indépendantes.

On n'exigera pas que les colonnes de la matrice de régression soient linéairement indépendantes.

A chaque ensemble de valeurs (8) correspond une valeur de la variable aléatoire y. Notons ces valeurs $\eta^1, \ldots, \eta^m$ et écrivons

$$\eta^1 = a_1^1 \theta^1 + \ldots + a_n^1 \theta^n + \varepsilon^1,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\eta^m = a_1^m \theta^1 + \ldots + a_n^m \theta^n + \varepsilon^m$$

ou sous forme matricielle

$$\boldsymbol{\eta} = A\boldsymbol{\theta} + \boldsymbol{\varepsilon}. \tag{9}$$

Selon (7), les variables aléatoires ε^i ont une espérance mathématique nulle. On les appelle *erreurs*, même si en réalité elles peuvent représenter des écarts naturels des valeurs de y par rapport à son espérance mathématique, écarts qui ne sont pas obligatoirement dus à des erreurs de mesure sur y.

Si ε^i sont des erreurs de mesure, il est naturel de supposer qu'elles ne sont pas corrélées entre elles et possèdent les mêmes variances, ce que nous admettrons dans la suite pour simplifier l'exposé.

S'agissant du système d'équations linéaires $\boldsymbol{\eta} = A\boldsymbol{\theta}$, la matrice-colonne $\boldsymbol{\varepsilon}$ est l'écart défini par la matrice-ligne $\boldsymbol{\theta} = \|\theta^1, \ldots, \theta^n\|$ des valeurs exactes du paramètre.

La pseudo-solution $\hat{\boldsymbol{\theta}}$ de ce système est appelée *estimation de $\boldsymbol{\theta}$ obtenue par la méthode des moindres carrés*. Elle vaut

$$\hat{\boldsymbol{\theta}} = A^+ \boldsymbol{\eta}. \tag{10}$$

Les estimations qui, telles que $\hat{\boldsymbol{\theta}}$, s'expriment linéairement par $\boldsymbol{\eta}$ s'appellent *estimations linéaires*.

L'estimation qui dépend d'un ensemble des variables aléatoires $\boldsymbol{\eta}$ est naturellement une variable aléatoire. L'estimation est dite *sans biais* ou *non biaisée* si son espérance mathématique est confondue avec l'ensemble des vraies valeurs du paramètre $\boldsymbol{\theta}$.

PROPOSITION 2. *Si les colonnes de la matrice de régression sont linéairement indépendantes, l'estimation $\hat{\boldsymbol{\theta}}$ obtenue par la méthode des moindres carrés est non biaisée.*

En effet, il découle des propriétés élémentaires de l'espérance mathématique que

$$\mathbf{E}(\hat{\boldsymbol{\theta}}) = A^+ \mathbf{E}(\boldsymbol{\eta}) = A^+ A\boldsymbol{\theta}, \tag{11}$$

ce qui est équivalent à l'égalité exigée

$$\mathsf{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta},$$

puisque les colonnes de la matrice A sont linéairement indépendantes par hypothèse.

Dans le cas général, la formule (11) ne permet d'affirmer l'absence de biais que pour quelques projections de $\hat{\boldsymbol{\theta}}$. Plus exactement, la matrice A^+A est idempotente, c'est-à-dire que $(A^+A)^2 = A^+A$ et, par suite, on déduit de (11) que

$$\mathsf{E}(A^+A\hat{\boldsymbol{\theta}}) = A^+A\boldsymbol{\theta}.$$

Selon la proposition 11 du § 1, $A^+A\hat{\boldsymbol{\theta}}$ est la projection orthogonale de $\hat{\boldsymbol{\theta}}$ sur le sous-espace $\mathscr{I}$ de $\mathscr{R}_n$. Ce sous-espace est composé des matrices-colonnes de la forme ${}^tA\boldsymbol{v}$ pour $\boldsymbol{v}$ quelconque. Donc si $\boldsymbol{p} \in \mathscr{I}$, on a $A^+A\boldsymbol{p} =$ $= \boldsymbol{p}$ et ${}^t\boldsymbol{p}A^+A = {}^t\boldsymbol{p}$. Par conséquent, pour tous les $\boldsymbol{p} \in \mathscr{I}$ on a

$$\mathsf{E}({}^t\boldsymbol{p}\hat{\boldsymbol{\theta}}) = {}^t\boldsymbol{p}\boldsymbol{\theta},$$

ce qui entraîne la

PROPOSITION 3. *Si les coefficients* $\boldsymbol{p}$ *appartiennent à* $\mathscr{I}$, *c'est-à-dire si* ${}^t\boldsymbol{p}A^+A = {}^t\boldsymbol{p}$, *alors* ${}^t\boldsymbol{p}\hat{\boldsymbol{\theta}}$ *est une estimation non biaisée de* ${}^t\boldsymbol{p}\boldsymbol{\theta}$.

Les fonctions linéaires sur $\mathscr{R}_m$, considéré comme l'ensemble de toutes les valeurs des paramètres, sont appelées fonctions paramétriques. La proposition 3 se rapporte en fait à l'estimation de la valeur d'une fonction paramétrique ${}^t\boldsymbol{p}\boldsymbol{\theta}$ sur l'ensemble des vraies valeurs du paramètre $\boldsymbol{\theta}$.

On a appelé plus haut estimation linéaire de la matrice-ligne des vraies valeurs du paramètre l'estimation qui dépend linéairement de $\boldsymbol{\eta}$, c'est-à-dire une fonction linéaire sur $\mathscr{R}_m$. Etudions les fonctions linéaires sur $\mathscr{R}_m$ qui sont des estimations linéaires non biaisées des fonctions paramétriques.

Par définition, la fonction ${}^t\boldsymbol{v}\boldsymbol{\zeta}$, où $\boldsymbol{\zeta} \in \mathscr{R}_m$, est *engendrée* par la fonction paramétrique $\psi(\boldsymbol{\theta}) = {}^t\boldsymbol{p}\boldsymbol{\theta}$ si pour tous les $\boldsymbol{\theta}$ on a

$${}^t\boldsymbol{v}A\boldsymbol{\theta} = {}^t\boldsymbol{p}\boldsymbol{\theta}, \tag{12}$$

ce qui équivaut à

$${}^t\boldsymbol{v}A = {}^t\boldsymbol{p},$$

ou

$${}^tA\boldsymbol{v} = \boldsymbol{p}. \tag{13}$$

Ainsi, la fonction paramétrique engendre une fonction sur $\mathscr{R}_m$ si et seulement si $\boldsymbol{p} \in \mathscr{I}$. Toutefois, cette condition n'implique pas en général que la fonction engendrée est unique. En effet, la solution générale du système d'équations (13) est

$$\boldsymbol{v}_c = ({}^tA)^+\boldsymbol{p} + (E_m - ({}^tA)^+({}^tA))\boldsymbol{c} = ({}^tA)^+\boldsymbol{p} + (E_m - AA^+)\boldsymbol{c},$$

où $\boldsymbol{c}$ est une matrice-colonne quelconque à m éléments (voir proposition 13, § 1).

Supposons que $f(\boldsymbol{\zeta}) = {}^t v\boldsymbol{\zeta}$ est une fonction linéaire sur $\mathscr{R}_m$ engendrée par une fonction paramétrique. Cherchons l'espérance mathématique et la variance de la valeur f sur la matrice-colonne $\boldsymbol{\eta}$ vérifiant (9). On admettra dans ce cas que les erreurs ε^i ne sont pas corrélées et présentent les mêmes variances $\mathbf{V}(\varepsilon^i) = \sigma^2$.

On a pour l'espérence mathématique

$$\mathbf{E}({}^t v\boldsymbol{\eta}) = \mathbf{E}({}^t vA\boldsymbol{\theta} + {}^t v\boldsymbol{\varepsilon}) = {}^t vA\boldsymbol{\theta} = {}^t p\boldsymbol{\theta}. \tag{14}$$

Donc, $\mathbf{E}({}^t v\boldsymbol{\eta})$ est égale à la valeur de la fonction génératrice sur l'ensemble des vraies valeurs du paramètre. De ce point de vue, ${}^t v\boldsymbol{\eta}$ *est une estimation non biaisée de* ${}^t p\boldsymbol{\theta}$. En appliquant ce résultat, on trouve

$$\mathbf{V}({}^t v\boldsymbol{\eta}) = \mathbf{E}({}^t v\boldsymbol{\eta} - {}^t p\boldsymbol{\theta})^2 = \mathbf{E}({}^t v(\boldsymbol{\eta} - A\boldsymbol{\theta}))^2 = \mathbf{E}({}^t v\boldsymbol{\varepsilon})^2.$$

Ensuite, utilisons l'égalité ${}^t v\boldsymbol{\varepsilon} = {}^t\boldsymbol{\varepsilon} v$ se vérifiant pour toutes matrices-colonnes :

$$\mathbf{E}({}^t v\boldsymbol{\varepsilon})^2 = \mathbf{E}({}^t v(\boldsymbol{\varepsilon}\,{}^t\boldsymbol{\varepsilon})v) = {}^t v\mathbf{E}(\boldsymbol{\varepsilon}\,{}^t\boldsymbol{\varepsilon})v = {}^t v\sigma^2 E_m v = \sigma^2\|v\|^2.$$

Ainsi donc,

$$\mathbf{V}({}^t v\boldsymbol{\eta}) = \sigma^2\|v\|^2. \tag{15}$$

Il en découle en particulier, qu'après avoir choisi la solution du système (13) de norme minimale, on choisit parmi toutes les fonctions engendrées par la fonction paramétrique donnée celle pour laquelle la variance $\mathbf{V}({}^t v\boldsymbol{\eta})$ est minimale. Le système (13) est supposé compatible. Il a une solution unique de norme minimale, à savoir la pseudo-solution

$$v_0 = ({}^t A)^+ p.$$

La valeur de la fonction de matrice-ligne de coefficients ${}^t v_0$ sur le vecteur $\boldsymbol{\eta}$ est selon (10) égale à

$${}^t v_0\boldsymbol{\eta} = {}^t pA^+\boldsymbol{\eta} = {}^t p\hat{\boldsymbol{\theta}},$$

c'est-à-dire à la valeur de la fonction génératrice sur l'estimation $\hat{\boldsymbol{\theta}}$ obtenue par la méthode des moindres carrés. Il s'ensuit le

THÉORÈME 1. *De toutes les estimations non biaisées de la forme* ${}^t v\boldsymbol{\eta}$ *d'une fonction paramétrique* ${}^t p\boldsymbol{\theta}$ *l'estimation* ${}^t v_0\boldsymbol{\eta} = {}^t p\hat{\boldsymbol{\theta}}$ *de la méthode des moindres carrés a la plus petite variance.*

En particulier, si les colonnes de la matrice A sont linéairement indépendantes, on a $\mathscr{I} = \mathscr{R}_m$, et toutes les fonctions paramétriques admettent une estimation. Il en est de même des fonctions $\boldsymbol{e}_i\boldsymbol{\theta}$, où $\boldsymbol{e}_i$ sont les lignes de la matrice unité. Ces fonctions sont égales à θ^i, et leurs estimations non biaisées de variance minimale seront les composantes $\hat{\theta}^i$ obtenues par la méthode des moindres carrés.

Le théorème démontré indique pourquoi il faut minimiser la somme des carrés, c'est-à-dire utiliser la norme euclidienne et non pas, disons, la somme des modules. Il faut remarquer que les avantages de la méthode des moindres carrés ne se sont pas révélés aussitôt. Ainsi, Laplace exigeait d'abord une minimisation de $\sum |\varepsilon^i|$.

Soit $\boldsymbol{\xi}$ un « vecteur aléatoire », c'est-à-dire une matrice-colonne composée de m variables aléatoires $\xi^1, \ldots, \xi^m$. Son espérance mathématique est une matrice-colonne $\boldsymbol{a}$ formée des éléments $a^i = \mathbf{E}(\xi^i)$. On appelle *matrice de covariance* de $\boldsymbol{\xi}$ la matrice

$$\mathbf{V}(\boldsymbol{\xi}) = \mathbf{E}[(\boldsymbol{\xi} - \boldsymbol{a})\,'(\boldsymbol{\xi} - \boldsymbol{a})].$$

Ses éléments diagonaux sont les variances des ξ^i correspondantes, tandis que les éléments non diagonaux, les covariances $\operatorname{cov}(\xi^i, \xi^j)$.

Voyons comment on peut estimer la matrice de covariance de l'estimation $\boldsymbol{\theta}$ obtenue par la méthode des moindres carrés si l'on suppose que les erreurs ε^i ne sont pas corrélées et possèdent une même variance égale à σ^2. Cela signifie que

$$\mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2 E_m.$$

De $\mathbf{V}(\boldsymbol{\eta}) = \mathbf{V}(\boldsymbol{\eta} - A\boldsymbol{\theta}) = \mathbf{V}(\boldsymbol{\varepsilon})$ il résulte que

$$\mathbf{V}(\boldsymbol{\eta}) = \sigma^2 E_m. \tag{16}$$

Il n'est pas difficile de vérifier que pour tout vecteur aléatoire $\boldsymbol{\xi}$ et toute matrice constante S il découle de $\boldsymbol{\xi} = S\boldsymbol{\xi}'$ que $\mathbf{E}(\boldsymbol{\xi}) = S\mathbf{E}(\boldsymbol{\xi}')$ et

$$\mathbf{V}(\boldsymbol{\xi}) = S\mathbf{V}(\boldsymbol{\xi}')({}'S).$$

Il s'ensuit en vertu de (10) que

$$\mathbf{V}\hat{\boldsymbol{\theta}} = A^+\sigma^2(E_m)\,'(A^+) = \sigma^2(A^+)\,'(A^+).$$

Il n'est pas difficile de vérifier que $(A^+)\,'(A^+) = ({}'AA)^+$. Donc,

$$\mathbf{V}\hat{\boldsymbol{\theta}} = \sigma^2({}'AA)^+.$$

On appelle *somme des carrés résiduelle* la grandeur $\|\hat{\boldsymbol{\varepsilon}}\|^2$ si $\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\eta} - A\hat{\boldsymbol{\theta}}$. Cherchons l'espérance mathématique de la somme des carrés résiduelle. On a

$$\begin{aligned}{}'\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}} = {}'(\boldsymbol{\eta} - AA^+\boldsymbol{\eta})(\boldsymbol{\eta} - AA^+\boldsymbol{\eta}) = {}'\boldsymbol{\eta}(E - 2AA^+ + {}'(A^+)\,'AAA^+)\boldsymbol{\eta} = \\ = {}'\boldsymbol{\eta}(E - AA^+)\boldsymbol{\eta}.\end{aligned}$$

Désignons les éléments de la matrice $E - AA^+$ par ρ_{ij}. Alors

$$'\boldsymbol{\eta}(E - AA^+)\boldsymbol{\eta} = \sum_{ij} \rho_{ij}\eta^i\eta^j,$$

et ${}^t\hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}$ peut être mis sous la forme

$$\sum_{i,j} \rho_{ij}(\eta^i\eta^j - \mathsf{E}(\eta^i)\mathsf{E}(\eta^j)) + \sum_{i,j} \rho_{ij}\mathsf{E}(\eta^i)\mathsf{E}(\eta^j). \qquad (17)$$

Calculons l'espérance mathématique de cette expression. Notons pour cela que selon la formule (16)

$$\mathsf{E}(\eta^i\eta^j - \mathsf{E}(\eta^i)\,\mathsf{E}(\eta^j)) = \operatorname{cov}(\eta^i, \eta^j) = \begin{cases} 0, & i \neq j, \\ \sigma^2, & i = j. \end{cases}$$

Le premier terme dans l'expression (17) est donc

$$\sigma^2 \sum_i \rho_{ii} = \sigma^2 \operatorname{tr}(E - AA^+).$$

Le second terme est une constante, et son espérance mathématique est donc égale à ce terme. Ecrit sous forme matricielle, ce terme est égal à

$${}^t(\mathsf{E}(\boldsymbol{\eta}))(E - AA^+)\,\mathsf{E}(\boldsymbol{\eta}),$$

ou ${}^t\boldsymbol{\theta}\,{}^tA(E - AA^+)A\boldsymbol{\theta}$. Or

$${}^t\boldsymbol{\theta}({}^tAA - {}^tAAA^+A)\boldsymbol{\theta} = 0.$$

Ainsi donc,

$$\mathsf{E}(\|\hat{\boldsymbol{\varepsilon}}\|^2) = \sigma^2 \operatorname{tr}(E - AA^+).$$

Pour trouver la trace de la matrice $E - AA^+$, souvenons-nous que la matrice AA^+ est idempotente (voir p. 466). Il en est donc de même de $E - AA^+$ et, par suite, ses nombres caractéristiques sont tous égaux à zéro ou à l'unité. Vu que $\operatorname{Rg}(E - AA^+) = m - r$, il y a exactement $m - r$ nombres caractéristiques égaux à l'unité. Donc,

$$\operatorname{tr}(E - AA^+) = m - r.$$

En définitive, on obtient

$$\mathsf{E}(\|\hat{\boldsymbol{\varepsilon}}\|^2) = \sigma^2(m - r).$$

Ce résultat est utilisé à la recherche de l'estimation du paramètre σ suivant la formule

$$\hat{\sigma}^2 = \frac{\|\hat{\boldsymbol{\varepsilon}}^2\|}{m - r}.$$

L'étude analytique de la régression peut être approfondie d'après le livre de Seber [35] ou tout autre cours de statistique mathématique. Les applications des matrices pseudo-inverses aux problèmes de statistique sont décrites dans le livre d'Albert [1].

CHAPITRE XV

SYSTÈMES D'INÉQUATIONS LINÉAIRES ET PROGRAMMATION LINÉAIRE

§ 1. Systèmes d'inéquations linéaires homogènes

1. Définitions fondamentales. Dans ce paragraphe, on étudiera les systèmes d'inéquations de la forme

$$\begin{gathered} a_1^1 x^1 + \ldots + a_n^1 x^n \geqslant 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_1^m x^1 + \ldots + a_n^m x^n \geqslant 0, \end{gathered} \tag{1}$$

où a_j^i sont des constantes réelles. Sans restreindre la généralité on peut poser que toutes les inéquations sont de même sens ($\geqslant 0$). En effet, on peut mettre l'inéquation $a_1 x^1 + \ldots + a_n x^n \leqslant 0$ sous la forme nécessaire en la multipliant par -1. De même, il est possible d'écrire sous la forme (1) tout système mixte composé d'équations et d'inéquations linéaires homogènes, vu qu'une équation linéaire homogène peut être écrite sous la forme d'un couple d'inéquations $a_1 x^1 + \ldots + a_n x^n \geqslant 0$ et $-a_1 x^1 - \ldots - a_n x^n \geqslant \geqslant 0$. D'autre part, l'absence dans le système étudié d'inégalités strictes (avec le signe $>$) constitue une hypothèse importante.

Convenons de poser que la matrice-colonne $\boldsymbol{x}$ est *positive*, et d'écrire $\boldsymbol{x} \geqslant \boldsymbol{0}$ si tous les éléments de la matrice-colonne sont positifs. Une même convention peut être faite pour les inégalités $\boldsymbol{x} \leqslant \boldsymbol{0}$, $\boldsymbol{x} \geqslant \boldsymbol{y}$, etc. Ceci étant, il faut avoir en vue que toutes les matrices-colonnes ne sont pas comparables, c'est-à-dire que les deux inégalités $\boldsymbol{x} > \boldsymbol{y}$ et $\boldsymbol{x} \leqslant \boldsymbol{y}$ peuvent ne pas être vraies.

En utilisant les notations introduites, on peut écrire le système d'inéquations (1) sous forme matricielle

$$A\boldsymbol{x} \geqslant \boldsymbol{0}. \tag{2}$$

Considérons une interprétation géométrique du système d'inéquations (1) en termes d'espaces vectoriels. Soit $\mathscr{L}_n$ un espace vectoriel réel rapporté à une base $\| e_1, \ldots, e_n \|$. L'ensemble des vecteurs dont les coordonnées vérifient l'inéquation linéaire homogène

$$a_1 x^1 + \ldots + a_n x^n \geqslant 0$$

est appelé *demi-espace fermé*. Si l'inégalité est stricte, le demi-espace est dit *ouvert*. Le demi-espace fermé est la réunion d'un demi-espace ouvert et d'un *sous-espace frontière* défini par l'équation

$$a_1 x^1 + \ldots + a_n x^n = 0.$$

On étudiera partout, sauf mention expresse du contraire, les demi-espaces fermés.

L'intersection d'un nombre fini de demi-espaces est appelée *cône polyédrique convexe fermé*. Vu que dans ce chapitre on ne rencontrera pas d'autres cônes, on laissera souvent tomber les qualificatifs « fermé », « convexe » et « polyédrique ». Selon cette définition, l'ensemble de tous les vecteurs dont les coordonnées vérifient le système d'inéquations linéaires homogènes est un cône polyédrique convexe.

Comme toutes les définitions où interviennent les coordonnées, les définitions du demi-espace et du cône dépendent de la base. On vérifie facilement qu'en réalité cette dépendance n'existe pas : un ensemble défini par le système (2) dans une base se définit par un système de même type dans toute autre base. En effet, soit $\boldsymbol{x} = S\boldsymbol{x}'$. Alors le système (2) est équivalent à $AS\boldsymbol{x}' \geqslant \mathbf{0}$.

En général, on n'aura pas besoin de procéder à des changements de base. La base étant fixée, on est en fait en présence d'un espace arithmétique n-dimensionnel. On notera le vecteur de la même façon que sa colonne de coordonnées, et ses composantes par la même lettre affectée d'un indice.

Donnons des exemples de cônes polyédriques dans l'espace tridimensionnel :

1) $x^1 \geqslant 0, \quad x^2 \geqslant 0, \quad x^3 \geqslant 0$ (octant positif) ;

2) $x^1 \geqslant 0, \quad x^2 \geqslant 0$ (dièdre) ;

3) $x^1 \geqslant 0, \quad -x^1 \geqslant 0$ (plan) ;

4) $x^1 \geqslant 0, \quad x^2 \geqslant 0, \quad x^3 \geqslant 0, \quad x^1 + x^2 - x^3 \geqslant 0$ (cône tétraédrique) ;

5) $x^1 \geqslant 0, \quad x^2 \geqslant 0, \quad x^3 \geqslant 0, \quad x^1 + x^2 - x^3 \geqslant 0, \quad -x^1 - x^2 + x^3 \geqslant 0$ (angle plan).

Laissons au soin du lecteur de montrer que sont des cônes : le vecteur nul, un sous-espace unidimensionnel et une *demi-droite*, c'est-à-dire un ensemble de vecteurs de la forme $\alpha \boldsymbol{x}$, où $\boldsymbol{x} \neq \mathbf{0}$ et $\alpha \geqslant 0$.

PROPOSITION 1. *Si les vecteurs $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ appartiennent à un cône polyédrique convexe $\mathscr{K}$, il en est de même des vecteurs $\boldsymbol{x}_1 + \boldsymbol{x}_2$ et $\alpha\boldsymbol{x}_1$ pour tout $\alpha \geqslant 0$.*

On le démontre facilement par substitution des coordonnées des vecteurs dans le système d'inéquations linéaires définissant le cône.

Soit donné un cône $\mathscr{K}$. On peut envisager son enveloppe linéaire, c'est-à-dire l'ensemble de toutes les combinaisons linéaires finies des vecteurs

appartenant au cône. L'enveloppe linéaire est le sous-espace de dimension minimale contenant le cône $\mathscr{K}$. La dimension de l'enveloppe linéaire du cône est appelée *dimension du cône.*

On dit que l'inéquation du système (1) représente une *contrainte-égalité* si pour toutes les solutions du système elle n'est vérifiée que comme une équation. En substituant les équations à de telles inéquations, on ne fera pas varier l'ensemble des solutions du système.

PROPOSITION 2. *Si p est le rang de la matrice composée des coefficients des contraintes-égalités, le cône $\mathscr{K}$ défini par le système* (1) *est contenu dans un sous-espace $(n - p)$-dimensionnel et y est défini par le système ne renfermant pas de contraintes-égalités.*

DÉMONSTRATION. Sans restreindre la généralité on peut supposer que les p premières inéquations du système représentent des contraintes-égalités et que les matrices-lignes de coefficients qui leur correspondent sont linéairement indépendantes. Admettons que ces matrices-lignes ont été complétées d'une façon quelconque jusqu'à une matrice carrée régulière S. Introduisons une nouvelle base liée à l'ancienne par la matrice de passage S^{-1}. Le cône étudié s'écrira dans la nouvelle base par un système d'inéquations linéaires

$$AS^{-1}y \geqslant 0. \tag{3}$$

Ce système ne diffère de celui de départ que par la transformation des variables et, par suite, aux contraintes-égalités du système initial correspondent des contraintes-égalités, et aux contraintes-inégalités, des contraintes-inégalités. Les p premières lignes de la matrice AS^{-1} se confondent avec les lignes de la matrice unité. Aussi les p premières inéquations du système (3) sont-elles de la forme $y^1 \geqslant 0, \dots, y^p \geqslant 0$. On peut les remplacer par les équations sans changer l'ensemble des solutions du système. Ceci étant, les autres contraintes-égalités sont satisfaites automatiquement. En tenant compte des égalités $y^1 = 0, \dots, y^p = 0$ dans les inéquations restantes, on obtient un système d'inéquations en $y^{p+1}, \dots, y^n$, qui est le système de contraintes-inégalités.

On voit donc que le cône est contenu dans le sous-espace engendré par les $n - p$-derniers vecteurs de base et est défini dans ce sous-espace par le système de contraintes-inégalités. La proposition est donc démontrée.

PROPOSITION 3. *Si le système d'inéquations linéaires homogènes* (1) *ne contient pas de contraintes-égalités, le système correspondant d'inéquations strictes*

$$\begin{array}{c} a_1^1 x^1 + \dots + a_n^1 x^n > 0, \\ \dots\dots\dots\dots\dots\dots\dots\dots \\ a_1^m x^1 + \dots + a_n^m x^n > 0 \end{array} \tag{4}$$

possède une solution.

DÉMONSTRATION. D'après la définition de la contrainte-inégalité, il existe pour la i-ième inéquation, $i = 1, \ldots, m$, une solution du système x_i telle que l'inégalité stricte soit vérifiée. Considérons la matrice-colonne $x_1 + \ldots + x_m$. Soit a^i la matrice-ligne de coefficients de la i-ième inéquation du système. En substituant la matrice-colonne $x_1 + \ldots + x_m$ dans cette inéquation, on obtient

$$a^i(x_1 + \ldots + x_m) = a^i x_1 + \ldots + a^i x_i + \ldots + a^i x_m.$$

Tous les termes sont ici positifs et le terme $a^i x_i$ est strictement positif. Ainsi, $x_1 + \ldots + x_m$ est la solution dont on démontre l'existence.

Soit $\mathcal{K}$ le cône défini par le système (1) ne contenant pas de contraintes-égalités. Les vecteurs dont les coordonnées vérifient le système (4) sont appelés *vecteurs intérieurs* au cône $\mathcal{K}$ et l'ensemble de tous les vecteurs intérieurs, son *intérieur*. En accord avec la proposition 2, tout cône dans un sous-espace est défini par un système de contraintes-inégalités. Le système correspondant d'inéquations strictes possède des solutions dans le sous-espace mentionné. L'ensemble de ces solutions est appelé *intérieur relatif du cône*.

PROPOSITION 4. *Chaque vecteur intérieur au cône $\mathcal{K}$ est contenu dans ce cône avec l'un de ses voisinages par rapport à une norme quelconque.*

Vu que toutes les normes sont équivalentes (théorème 1, § 3, ch. XI), il suffit de démontrer l'assertion pour la c-norme

$$\|x\|_c = \max_i |x^i|.$$

Considérons d'abord une seule inéquation et un vecteur x_0 pour lequel

$$a_1 x_0^1 + \ldots + a_n x_0^n = h > 0.$$

Posons

$$\varepsilon = h \left(\sum_{i=1}^{n} |a_i| \right)^{-1}. \tag{5}$$

(Le nombre ε est bien défini, vu que tous les coefficients $a_1, \ldots, a_n$ ne sont pas nuls : l'inéquation dont tous les coefficients sont nuls est une contrainte-égalité.) Si le vecteur x est tel que $\|x - x_0\|_c < \varepsilon$, il vérifie aussi l'inéquation stricte. En effet, en désignant $x_0^i - x^i$ par δ^i, on obtient

$$\sum_{i=1}^{n} a_i x^i = \sum_{i=1}^{n} a_i x_0^i - \sum_{i=1}^{n} a_i \delta^i \geqslant h - \left| \sum_{i=1}^{n} a_i \delta^i \right|.$$

Or

$$\left| \sum_{i=1}^{n} a_i \delta^i \right| \geqslant \sum_{i=1}^{n} |\delta_i| |a_i| \leqslant \max_i |\delta_i| \sum_{i=1}^{n} |a_i| < h.$$

D'où on conclut que

$$\sum_{i=1}^{n} a_i x^i > 0.$$

Considérons maintenant le système d'inéquations (4) et le vecteur x_0 qui est la solution de ce système. Chacune des inéquations du système définit, d'après la formule (5), le rayon d'un voisinage du vecteur x_0, dont tout vecteur vérifie cette inéquation. Le plus petit de ces voisinages est composé de vecteurs vérifiant le système (4).

Il est maintenant aisé de démontrer la

PROPOSITION 5. *Si le cône $\mathscr{K}$ dans l'espace $\mathscr{L}_n$ est défini par le système d'inéquations linéaires sans contraintes-égalités, la dimension de $\mathscr{K}$ vaut la dimension de l'espace $\mathscr{L}_n$.*

Pour démontrer cette proposition, considérons un vecteur intérieur quelconque x_0 du cône $\mathscr{K}$. Supposons que ses coordonnées sont $x_0^1, \ldots, x_0^n$, et ε est le rayon d'un voisinage du vecteur x_0 composé de vecteurs intérieurs à $\mathscr{K}$. Considérons les vecteurs de coordonnées

$$\begin{array}{llll} x_0^1 - \eta, & x_0^2, & \ldots, & x_0^n, \\ x_0^1, & x_0^2 - \eta, & \ldots, & x_0^n, \\ \multicolumn{4}{c}{\dotfill} \\ x_0^1, & x_0^2, & \ldots, & x_0^n - \eta, \end{array}$$

où $|\eta| < \varepsilon$ et η est différent des nombres caractéristiques de la matrice dont toutes les lignes coïncident avec la ligne $\| x_0^1, \ldots, x_0^n \|$. On voit aisément que ces vecteurs sont linéairement indépendants et appartiennent à $\mathscr{K}$. La proposition est démontrée.

Il ressort de ce qui précède la

PROPOSITION 6. *La dimension du cône $\mathscr{K}$ défini par le système d'inéquations linéaires* (1) *vaut $n - p$, où p est le rang de la matrice composée des coefficients de toutes les contraintes-égalités du système.*

2. Structure du cône polyédrique convexe. Soit $\mathscr{K}$ le cône défini par le système d'inéquations (1). Considérons un sous-ensemble du cône $\mathscr{K}$ obtenu par substitution de certaines contraintes-inégalités du système (1) par des équations correspondantes. On voit aussitôt que ce sous-ensemble de $\mathscr{K}$ est un cône de dimension inférieure. On l'appelle *face* du cône $\mathscr{K}$.

Un cas particulier de face est le sous-espace défini par le système d'équations linéaires

$$\begin{array}{c} a_1^1 x^1 + \ldots + a_n^1 x^n = 0, \\ \dotfill \\ a_1^m x^1 + \ldots + a_n^m x^n = 0. \end{array} \qquad (6)$$

Cette face porte le nom de *face minimale* du cône. Cette appellation est liée au fait que toute face $\mathscr{K}'$ renferme la face minimale. La dimension de la face minimale est $n - r$, où r est le rang de la matrice A composée des coefficients du système (1).

En considérant la face $\mathscr{K}'$ du cône $\mathscr{K}$ comme un nouveau cône, on est en mesure de définir les faces de $\mathscr{K}'$. Il s'avère qu'elles sont des faces du cône $\mathscr{K}$. La face minimale du cône $\mathscr{K}$ est encore la face minimale de l'une quelconque de ses faces.

La face minimale du cône peut s'avérer le sous-espace nul. Dans ce cas, le cône est dit *pointé*. Pour qu'un cône soit pointé il faut et il suffit que $\mathrm{Rg}\, A = n$. Si $\mathrm{Rg}\, A < n$, le cône est dit *épointé*.

Introduisons la définition suivante. Soient $\mathscr{P}_1, \ldots, \mathscr{P}_q$ des ensembles de vecteurs de l'espace vectoriel $\mathscr{L}_n$. On appellera *somme* $\mathscr{P}_1 + \ldots + \mathscr{P}_q$ de ces ensembles l'ensemble de tous les vecteurs de la forme $\boldsymbol{x} = \boldsymbol{x}_1 + \ldots$ $\ldots + \boldsymbol{x}_q$, où $\boldsymbol{x}_i \in \mathscr{P}_i$, $i = 1, \ldots, q$.

Notons que la somme usuelle des sous-espaces est leur somme au sens indiqué plus haut.

Maintenant on peut formuler et démontrer la proposition qui suit.

PROPOSITION 7. *Tout cône polyédrique convexe $\mathscr{K}$ est la somme d'un cône pointé $\mathscr{K}^1$ et de sa face minimale $\mathscr{L}^0$. Chaque face $\mathscr{K}'$ du cône $\mathscr{K}$ est la somme de $\mathscr{L}^0$ et d'une face du cône $\mathscr{K}^1$. Inversement, chacune de ces sommes est une face du cône $\mathscr{K}$.*

DÉMONSTRATION. Soit $\mathscr{L}^1$ un sous-espace tel que $\mathscr{L}_n = \mathscr{L}^0 + \mathscr{L}^1$ et $\mathscr{L}^0 \cap \mathscr{L}^1 = 0$. Considérons l'ensemble de vecteurs $\mathscr{K}^1 = \mathscr{L}^1 \cap \mathscr{K}$. Il est défini par le système d'inéquations obtenu par réunion du système d'inéquations (1) du cône $\mathscr{K}$ et du système d'équations du sous-espace $\mathscr{L}^1$. Donc, $\mathscr{K}^1$ est un cône. La face minimale de $\mathscr{K}^1$ est définie par le système d'équations obtenu par substitution des équations à toutes les inéquations du système (1) et par adjonction des équations du sous-espace $\mathscr{L}^1$. Or c'est le système d'équations de $\mathscr{L}^0 \cap \mathscr{L}^1$ qui ne possède que la solution triviale. Donc, le cône $\mathscr{K}^1$ est pointé.

Ensuite, pour tout $\boldsymbol{x} \in \mathscr{L}_n$ on a la décomposition $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{x}_1$, où $\boldsymbol{x}_0 \in \mathscr{L}^0$ et $\boldsymbol{x}_1 \in \mathscr{L}^1$. Si $\boldsymbol{x} \in \mathscr{K}$, le vecteur $\boldsymbol{x}_1$ qui est la somme des vecteurs $\boldsymbol{x}$ et $-\boldsymbol{x}_0$ de $\mathscr{K}$ appartient aussi à $\mathscr{K}$. Ainsi, $\boldsymbol{x}_1 \in \mathscr{K}^1$. Inversement, si $\boldsymbol{x}_0 \in \mathscr{L}_0$ et $\boldsymbol{x}_1 \in \mathscr{K}^1$, on a $\boldsymbol{x}_0 + \boldsymbol{x}_1 \in \mathscr{K}$. Ceci achève la démonstration de la première assertion.

Pour démontrer la seconde assertion, il suffit de remarquer que l'intersection $\mathscr{K}' \cap \mathscr{L}^1$ est une face du cône $\mathscr{K}^1$. Or ce fait est évident parce que le système d'inéquations de $\mathscr{K}' \cap \mathscr{L}^1$ se déduit du système de $\mathscr{K}^1 = \mathscr{K} \cap \mathscr{L}^1$ par substitution d'équations à quelques inéquations, plus précisement à celles qui deviennent équations sur la face $\mathscr{K}'$ du cône $\mathscr{K}$.

L'assertion inverse se démontre de façon aussi bien simple.

Une conclusion importante sur la structure des cônes polyédriques peut être tirée de la proposition suivante.

PROPOSITION 8. *Soit $\mathscr{K}$ un cône polyédrique convexe pointé de dimension $\geqslant 2$. Chaque vecteur $x_0 \in \mathscr{K}$ peut être représenté comme une somme de deux vecteurs appartenant aux faces de $\mathscr{K}$.*

DÉMONSTRATION. Pour le vecteur appartenant à une face l'assertion est évidente. Soit x_0 un vecteur n'appartenant à aucune face.

Désignons par $a^1, \ldots, a^m$ les lignes de la matrice du système d'inéquations définissant $\mathscr{K}$ et considérons un vecteur x_1 de $\mathscr{K}$ non colinéaire au vecteur x_0. On supposera qu'aux contraintes-inégalités sont associés les numéros $i \leqslant s$. On a alors pour ces i les inégalités strictes $a^i x_0 > 0$.

Considérons les nombres

$$\lambda_i = \frac{a^i x_1}{a^i x_0}, \quad i = 1, \ldots, s,$$

et démontrons que parmi ces derniers deux au moins sont différents. En effet, s'il existe un nombre λ tel que pour tous les $i \leqslant s$

$$\lambda(a^i x_0) - a^i x_1 = 0,$$

le vecteur $\lambda x_0 - x_1$ vérifie toutes les contraintes-inégalités en tant que racine des équations correspondantes. En outre, pour tous les $i > s$, on a $a^i x_0 = 0$ et $a^i x_1 = 0$. Il s'ensuit que $\lambda x_0 - x_1$ vérifie également toutes les contraintes-égalités. Cela signifie que $\lambda x_0 - x_1$ appartient à la face minimale du cône $\mathscr{K}$ et, par suite, est nul car le cône est pointé. Dans ce cas, le vecteur x_1 est colinéaire à x_0, ce qui contredit l'hypothèse.

Ainsi donc, parmi les nombres λ_i il existe un maximal λ_l et un minimal λ_j. Pour λ_l, on a

$$\lambda_l(a^i x_0) - (a^i x_1) \geqslant 0, \quad i \neq l, i \leqslant s,$$
$$\lambda_l(a^l x_0) - (a^l x_1) = 0.$$

De plus,

$$\lambda_l(a^i x_0) - (a^i x_1) = 0, \quad i > s,$$

vu que x_0 et x_1 vérifient les contraintes-égalités. Toutes ces relations signifient que le vecteur $y = \lambda_l x_0 - x_1$ appartient à une face du cône $\mathscr{K}$. On démontre de façon analogue que le vecteur $z = x_1 - \lambda_j x_0$ appartient à une autre face du cône $\mathscr{K}$. Maintenant l'assertion nécessaire découle facilement de l'égalité $y + z = (\lambda_l - \lambda_j)x_0$.

Les faces d'un cône polyédrique convexe possèdent la propriété suivante.

PROPOSITION 9. *Si le vecteur x de la face $\mathscr{K}'$ est une somme de plusieurs vecteurs du cône, tous ces vecteurs appartiennent à $\mathscr{K}'$.*

En effet, soit a^i la matrice-ligne des coefficients d'une des inéquations du système, qui deviennent équations sur la face $\mathscr{K}'$, et soit $x = x_1 + \ldots + x_l$ la décomposition dont il s'agit. Alors,

$$a^i x = a^i x_1 + \ldots + a^i x_l = 0.$$

Or la somme de nombres positifs ne s'annule que si tous ces nombres sont égaux à zéro. Donc $a^i x_j = 0$ pour tous les $j = 1, \ldots, l$, d'où la proposition.

On obtient les faces unidimensionnelles d'un cône par substitution des équations à $n - 1$ inéquations linéairement indépendantes du système définissant le cône. S'il reste encore une contrainte-inégalité indépendante, la face unidimensionnelle est une demi-droite. Les faces unidimensionnelles sont appelées *arêtes* du cône polyédrique. Dans un cône pointé, la face unidimensionnelle est obligatoirement une demi-droite.

Proposition 10. *Un cône pointé est la somme de ses arêtes.*

Démontrons d'abord que chaque vecteur x d'un cône pointé peut être représenté sous la forme

$$x = \alpha_1 x_1 + \ldots + \alpha_N x_N,$$

où tous les $\alpha_i \geqslant 0$, et x_i sont des vecteurs non nuls appartenant aux arêtes du cône.

La démonstration sera donnée par récurrence sur la dimension du cône. Dans un cône pointé bidimensionnel, les arêtes constituent des faces et, par suite, la démonstration de la proposition 10 coïncide avec celle de la proposition 8.

Pour un cône de dimension n, chaque vecteur x peut, en vertu de la même proposition 8, être décomposé en somme $x = x_1 + x_2$, où x_1 et x_2 appartiennent aux faces. Les faces sont des cônes pointés de dimension $< n$ et selon l'hypothèse de récurrence on a les décompositions

$$x_1 = \alpha_1 y_1 + \ldots + \alpha_K y_K$$

et

$$x_2 = \beta_1 z_1 + \ldots + \beta_L z_L,$$

où tous les $\alpha_i, \beta_j \geqslant 0$, et y_i et z_j sont des vecteurs non nuls appartenant aux arêtes. D'où on obtient la proposition nécessaire.

L'assertion inverse est évidente. En effet, si $x_1, \ldots, x_N$ sont les vecteurs directeurs des arêtes, toute combinaison linéaire positive de ces vecteurs appartient au cône selon la proposition 1.

Pour passer des cônes pointés aux cônes de forme générale, démontrons la

Proposition 11. *Le sous-espace $\mathscr{L}_q$ de dimension q est la somme de $q + 1$ demi-droites.*

DÉMONSTRATION. Soit $\|e_1, \ldots, e_q\|$ une base dans $\mathscr{L}_q$. Posons $f = -(e_1 + \ldots + e_q)$. Alors pour tout $i = 1, \ldots, q$ on a l'égalité

$$-e_i = f + \sum_{j \neq i} e_j .$$

Un vecteur quelconque x de $\mathscr{L}_q$ admet la décomposition $x = \xi^1 e_1 + \ldots + \xi^q e_q$. Si $\xi^i < 0$ pour un i quelconque, remplaçons $\xi^i e_i$ dans la décomposition de x par

$$|\xi^i| \left(f + \sum_{j \neq i} e_j\right).$$

Après la réduction des termes semblables on obtient la décomposition de x suivant les vecteurs $e_1, \ldots, e_q, f$ à coefficients positifs.

Inversement, on voit aussitôt que toute combinaison linéaire de ces vecteurs à coefficients positifs appartient à $\mathscr{L}_q$. La proposition est démontrée.

Les propositions 7, 10 et 11 permettent d'énoncer le théorème suivant.

THÉORÈME 1. *Tout cône polyédrique convexe fermé est la somme d'un nombre fini de demi-droites ou, ce qui revient au même, l'ensemble des combinaisons linéaires d'un nombre fini de vecteurs à coefficients positifs.*

Il se peut que pour certains objectifs la représentation se rattachant à la proposition 7 soit plus commode : tout vecteur du cône est la somme d'une combinaison linéaire de vecteurs de base de la face minimale et d'une combinaison linéaire positive de vecteurs non nuls appartenant aux arêtes d'un cône pointé.

Le théorème 1 peut être formulé de façon différente :

THÉORÈME 1A. *La solution générale d'un système d'inéquations linéaires homogènes peut être écrite sous la forme*

$$x = \alpha_1 x_1 + \ldots + \alpha_N x_N,$$

où $x_1, \ldots, x_N$ est un ensemble de solutions et les coefficients $\alpha_1, \ldots, \alpha_N$ sont positifs.

On se servira de la terminologie suivante. L'ensemble fini de demi-droites dont la somme est un cône, sera appelé *système de génératrices* du cône et l'on dira qu'elles *engendrent* le cône. Quant aux vecteurs directeurs de ces demi-droites, on dira également qu'ils engendrent le cône.

On appelle *carcasse* du cône le système minimal (en quantité) de ses génératrices. Par abus de langage, on appellera aussi carcasse l'ensemble des vecteurs directeurs des demi-droites formant la carcasse du cône.

S'agissant du système d'inéquations linéaires, on dira que l'ensemble des solutions mentionné dans le théorème 1a est une *famille complète de*

solutions. On appellera *famille fondamentale* la plus petite (en quantité) famille complète de solutions.

Notons qu'il existe un procédé laborieux qui permet en principe d'obtenir toutes les arêtes du cône pointé : il suffit de passer en revue tous les sous-systèmes de rang $n - 1$ du système d'équations linéaires (6). (Ces systèmes existent car le rang du système (6) vaut n et l'élimination d'une équation diminue le rang au plus de 1.) Chacun de ces systèmes possède une famille fondamentale de solutions formée d'une solution unique. Si cette solution $\boldsymbol{x}$ vérifie le système (1), elle définit l'arête du cône ; si $\boldsymbol{x}$ vérifie le système (1) de signes contraires ($\leqslant$), c'est $-\boldsymbol{x}$ qui définit l'arête du cône. Dans les autres cas, $\boldsymbol{x}$ ne présente pas d'intérêt.

En parlant des familles fondamentales de solutions des systèmes d'inéquations linéaires homogènes, il faut avoir en vue qu'elles diffèrent essentiellement des familles fondamentales de solutions des systèmes d'équations homogènes. Plus précisément, la décomposition de la solution du système d'inéquations suivant la famille fondamentale de solutions n'est en général pas unique. Donnons un exemple correspondant.

Supposons que le cône tétraédrique est défini dans un espace tridimensionnel par le système d'inéquations

$$x^1 \geqslant 0, \quad x^2 \geqslant 0, \quad x^3 \geqslant 0, \quad x^1 + x^2 - x^3 \geqslant 0.$$

On voit aussitôt que la carcasse de ce cône peut, par exemple, être composée des vecteurs $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{f} = \boldsymbol{e}_1 + \boldsymbol{e}_3$ et $\boldsymbol{g} = \boldsymbol{e}_2 + \boldsymbol{e}_3$, où $\boldsymbol{e}_1$, $\boldsymbol{e}_2$ et $\boldsymbol{e}_3$ sont les vecteurs de base. On remarque de même aisément que le vecteur $\boldsymbol{e}_2 + \boldsymbol{f}$ appartenant au cône peut aussi être représenté sous la forme $\boldsymbol{e}_1 + \boldsymbol{g}$.

Pour décrire la carcasse d'un cône quelconque, introduisons la définition suivante. On appelle *vecteur extrémal* du cône $\mathscr{K}$ un vecteur $\boldsymbol{x} \in \mathscr{K}$ qui n'admet aucune représentation de forme $\boldsymbol{x}_1 + \boldsymbol{x}_2$, où $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ sont des vecteurs non colinéaires appartenant à $\mathscr{K}$. Autrement dit, le vecteur $\boldsymbol{x}$ est extrémal si de $\boldsymbol{x} = \boldsymbol{x}_1 + \boldsymbol{x}_2$ et $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathscr{K}$ il ressort que $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ sont colinéaires.

Il est évident que le vecteur $\lambda\boldsymbol{x}$ proportionnel au vecteur extrémal $\boldsymbol{x}$ pour un λ positif est aussi extrémal. La demi-droite composée de vecteurs extrémaux est dite *extrémale*.

PROPOSITION 12. *Si le cône $\mathscr{K}$ est pointé, ses arêtes sont ses seules demi-droites extrémales.*

En effet, si le vecteur $\boldsymbol{x} \in \mathscr{K}$ n'appartient pas à l'arête de $\mathscr{K}$, il appartient soit à l'intérieur de $\mathscr{K}$, soit à l'intérieur de sa face de dimension $\geqslant 2$. Dans ce cas, il n'est plus extrémal car se décompose en somme de deux vecteurs appartenant aux faces du cône auquel il est intérieur. Ainsi, seules les arêtes peuvent être des demi-droites extrémales. Or elles sont justement des demi-droites extrémales, ce qui découle de la proposition 9. La proposition est démontrée.

Chaque demi-droite extrémale du cône est contenue obligatoirement dans sa carcasse, car les vecteurs de cette demi-droite ne sont pas des combinaisons linéaires positives de vecteurs d'autres demi-droites. On peut donc énoncer le corollaire immédiat des propositions 10 et 12.

PROPOSITION 13. *La carcasse d'un cône pointé ne contient que ses arêtes.*

Démontrons maintenant la proposition qui suit.

PROPOSITION 14. *La carcasse d'un cône $\mathscr{K}$ est la réunion des carcasses de sa face minimale $\mathscr{L}_0$ et du cône pointé $\mathscr{K}_1$ tel que $\mathscr{K} = \mathscr{L}_0 + \mathscr{K}_1$.*

DÉMONSTRATION. Selon la proposition 7, le cône $\mathscr{K}$ possède des faces de dimension $n - r + 1$ dont chacune est la somme de la face minimale $\mathscr{L}_0$ du cône $\mathscr{K}$ et d'une des arêtes du cône $\mathscr{K}_1$. Soit $\mathscr{M}$ une de ces faces. On obtient la carcasse de $\mathscr{M}$ si à la carcasse de $\mathscr{L}_0$ on joint le vecteur directeur de l'arête du cône $\mathscr{K}_1$ qui est contenue dans $\mathscr{M}$. En effet, il va de soi que tous les vecteurs de $\mathscr{M}$ se décomposent en des combinaisons linéaires positives de ce système de $n - r + 2$ vecteurs. La carcasse de $\mathscr{M}$ ne peut posséder un nombre inférieur de vecteurs, car l'ensemble des combinaisons linéaires positives de $n - r + 1$ vecteurs est dans l'espace $(n - r + 1)$-dimensionnel soit un cône pointé (si ces vecteurs constituent une base), soit un cône de moindre dimension (si les vecteurs sont linéairement dépendants).

Ensuite, en vertu de la proposition 9, la carcasse du cône $\mathscr{K}$ doit contenir les carcasses de toutes ses faces $(n - r + 1)$-dimensionnelles. Le système minimal répondant à cette exigence est le système de vecteurs dont il s'agit dans l'énoncé de la proposition, ce qui achève la démonstration.

La proposition démontrée peut être considérée comme une précision du théorème 1, permettant de décrire le système minimal de vecteurs qui engendre le cône.

On démontrera plus bas le théorème inverse du théorème 1, mais avant il nous faut étudier les inéquations linéaires qui se déduisent du système donné d'inéquations linéaires.

3. Inéquations résultant d'un système d'inéquations linéaires. Considérons une combinaison linéaire positive des inéquations du système (1). Cette inéquation linéaire est de la même forme que les inéquations du système. La matrice-ligne de ses coefficients est $\boldsymbol{uA}$, où $\boldsymbol{u}$ est la matrice-ligne $\| u_1, \ldots, u_m \|$ d'éléments positifs et A la matrice du système.

Il est évident que l'inéquation $\boldsymbol{uAx} \geqslant \boldsymbol{0}$ est une conséquence du système $\boldsymbol{Ax} \geqslant \boldsymbol{0}$ au sens qu'elle se vérifie pour toutes ses solutions.

On obtiendra ici le résultat fondamental consistant dans le fait que chaque inéquation linéaire homogène se déduisant du système $\boldsymbol{Ax} \geqslant \boldsymbol{0}$ est de la forme $\boldsymbol{uAx} \geqslant \boldsymbol{0}$ pour $\boldsymbol{u} \geqslant \boldsymbol{0}$. Ce résultat est connu sous le nom de *théorème de Farkas*. Démontrons préalablement la

PROPOSITION 15. *Les systèmes* $Ax \geqslant 0$ *et* $uA = 0$, $u \geqslant 0$ *possèdent les solutions* x_0 *et* u_0 *pour lesquelles*

$$Ax_0 + {}^t u_0 > 0. \tag{7}$$

Démontrons d'abord qu'il existe des solutions pour lesquelles la première composante du premier membre de (7) est strictement positive. Après quoi la démonstration se fera sans difficulté.

Si $m = 1$, c'est-à-dire si la matrice A est composée d'une seule ligne, l'assertion est évidente. En effet, si l'unique ligne a^1 est nulle, posons $x = 0$ et $u = 1$. Dans le cas contraire, posons $x = {}^t(a^1)$ et $u = 0$.

Admettons maintenant que l'assertion est démontrée pour les matrices composées de $m - 1$ lignes et démontrons-la pour une matrice A à m lignes $a^1, \ldots, a^m$. Les lignes $a^1, \ldots, a^{m-1}$ constituent la matrice $\tilde{A}$ à laquelle on peut appliquer l'hypothèse de récurrence. Il existe donc une matrice-colonne x à n éléments et une matrice-ligne positive $\| u^1, \ldots, u^{m-1} \|$ pour lesquelles

$$Ax \geqslant 0, \quad u\tilde{A} = 0, \quad a^1 x + u^1 > 0.$$

Si $a^m x \geqslant 0$, alors x et $\| u^1, \ldots, u^{m-1}, 0 \|$ sont les solutions cherchées, et l'assertion est démontrée. Admettons que $a^m x < 0$.

Composons la matrice B de type $(m - 1, n)$ à partir des lignes

$$b^i = a^i - \alpha^i a^m, \quad i = 1, \ldots, m - 1,$$

où

$$\alpha_i = \frac{a^i x}{a^m x}.$$

Il est évident que la matrice-colonne Bx est composée des éléments $a^i x + \alpha^i a^m x$ et, par suite, est nulle. Appliquons à la matrice B l'hypothèse de récurrence. On obtient la matrice-colonne y et la matrice-ligne v telles que

$$By \geqslant 0, \quad vB = 0, \quad v \geqslant 0,$$
$$b^1 y + v^1 > 0.$$

Considérons la matrice-ligne

$$w = \| v^1, \ldots, v^{m-1}, - \sum_{i=1}^{m-1} \alpha_i v^i \|.$$

Elle est positive car tous les $v^i \geqslant 0$, $\alpha_i \leqslant 0$ en vertu de $a^i x \geqslant 0$ et $a^m x < 0$. On voit aussitôt que

$$wA = vB = 0.$$

Considérons la matrice-colonne $z = y + \beta x$, où β est choisi de manière que $a^m z = 0$. Il faut pour cela que $\beta = -a^m y / a^m x$. On a

$$Az = \left\| \begin{matrix} \bar{A}z \\ 0 \end{matrix} \right\| = \left\| \begin{matrix} Bz \\ 0 \end{matrix} \right\| = \left\| \begin{matrix} By \\ 0 \end{matrix} \right\| \leqslant 0. \tag{8}$$

On a utilisé ici l'égalité $Bx = 0$ démontrée plus haut.

Ensuite, il résulte de (8) que $a^1 z = b^1 y$. En outre, $w^1 = v^1$. Donc, $a^1 z + w^1 = b^1 y + v^1 > 0$ et, par suite, z et w sont la matrice-colonne et la matrice-ligne cherchées pour la matrice A.

Démontrons maintenant qu'il est possible de trouver une matrice-colonne x_0 et une matrice-ligne u_0 telles que $Ax_0 \geqslant 0$, $u_0 A = 0$, $u_0 \geqslant 0$ et $Ax_0 + {}^t u_0 > 0$. Pour cela appliquons l'assertion démontrée plus haut à la matrice $P_i A$, où P_i est la matrice de permutation qui met la i-ième ligne à la première place. On obtient la matrice-ligne w_i et la matrice-colonne z_i pour lesquelles la i-ième composante du premier membre de (7) est strictement positive. Considérons

$$x_0 = \sum_{i=1}^{m} z_i \quad \text{et} \quad u_0 = \sum_{i=1}^{m} w_i.$$

Il est aisé de montrer que x_0 et u_0 vérifient toutes les relations nécessaires, ce qui achève la démonstration.

Démontrons maintenant le théorème de Farkas formulé de la façon suivante.

THÉORÈME 2. *Quelles que soient la matrice A à m lignes et n colonnes et la matrice-ligne b à n éléments, un et un seul des deux systèmes*

$$Ax \geqslant 0, \quad bx < 0$$

et

$$uA = b, \quad u \geqslant 0$$

est résoluble.

Remarquons tout d'abord que cette assertion coïncide en effet avec celle qui a été formulée au début de ce point : si le premier système est incompatible, $Ax \geqslant 0$ entraîne $bx \geqslant 0$, de sorte que le second système est compatible, c'est-à-dire que b est une combinaison linéaire positive des lignes de A. Inversement, si le premier système est compatible, $bx \geqslant 0$ ne se déduit pas de $Ax \geqslant 0$, si bien que le second système est incompatible, c'est-à-dire que b n'est pas une combinaison linéaire positive des lignes de A.

Pour démontrer le théorème, considérons la matrice $\bar{A}$ déduite de A par adjonction de la ligne $-b$. En vertu de la proposition 15, il existe une

matrice-colonne x à n éléments et une matrice-ligne positive u à $m + 1$ éléments, telles que

$$\bar{A}x \geqslant \mathbf{0}, \quad \bar{u}A = \mathbf{0}, \quad \bar{A}x + {}^t\bar{u} > \mathbf{0}.$$

En dégageant le dernier élément de la matrice-ligne $\bar{u}$

$$\bar{u} = \| u_1, \ldots, u_m, u_{m+1} \| = \| u, u_{m+1} \|,$$

on obtient les relations suivantes :

$$Ax \geqslant \mathbf{0}, \quad -bx \geqslant 0, \qquad uA - u_{m+1}b = \mathbf{0}.$$

Dégageons la dernière composante de la matrice-colonne $Ax + {}^t\bar{u}$:

$$-bx + u_{m+1} > 0.$$

Cela veut dire que : soit $u_{m+1} > 0$, soit $bx < 0$. Dans le premier cas, est compatible le second système : $u^0A = b$, où

$$u_i^0 = \frac{u_i}{u_{m+1}} \geqslant 0, \quad i = 1, \ldots, m.$$

Dans le deuxième, est compatible le premier système : $Ax \geqslant \mathbf{0}$, $bx < 0$.

Démontrons que les deux systèmes ne peuvent à la fois être compatibles. En effet, il découle des quatre relations que $uAx \geqslant 0$ et $uAx = bx < 0$. Le théorème est démontré.

Si on remplace la matrice A par la matrice transposée ${}^tA$, les matrices-lignes par les matrices-colonnes, et *vice versa,* on obtient la formulation équivalente suivante du théorème.

COROLLAIRE. *Quelles que soient la matrice A et la matrice-colonne b, un et un seul des deux systèmes*

$$Ax = b, \quad x \geqslant \mathbf{0}$$

et

$$uA \geqslant \mathbf{0}, \quad ub < 0.$$

est résoluble.

Ce corollaire représente la condition d'existence de solutions positives pour le système d'équations linéaires. Il est utile de le comparer au théorème de Fredholm qu'on peut formuler de la façon suivante :

Pour toute matrice A et toute matrice-colonne b un et un seul des deux systèmes

$$Ax = b$$

et

$$uA = \mathbf{0}, \quad ub < 0$$

est résoluble. (On laisse au soin du lecteur de démontrer que l'assertion donnée est équivalente au théorème de Fredholm.)

On est maintenant en mesure de démontrer le théorème inverse du théorème 1. Le théorème 1 et son inverse montrent que le cône polyédrique convexe peut être défini non seulement comme l'intersection d'une famille finie de sous-espaces mais aussi comme l'ensemble de toutes les combinaisons linéaires positives d'un nombre fini de vecteurs.

THÉORÈME 3. *Soit $\mathscr{M}$ l'ensemble de toutes les combinaisons linéaires positives des vecteurs $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$. Alors $\mathscr{M}$ est l'intersection d'une famille finie de sous-espaces.*

Soit A une matrice a n lignes et m colonnes $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_m$. Le vecteur $\boldsymbol{b}$ appartient à $\mathscr{M}$ si et seulement si $\boldsymbol{b} = A\boldsymbol{x}, \boldsymbol{x} \geqslant \boldsymbol{0}$. En vertu du corollaire du théorème 2 cette condition est équivalente au fait que l'inéquation $\boldsymbol{ub} \geqslant 0$ découle du système d'inéquations $\boldsymbol{u}A \geqslant \boldsymbol{0}$ ou, ce qui revient au même, du système ${}^tA{}^t\boldsymbol{u} \geqslant \boldsymbol{0}$.

Notons $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ la famille fondamentale de solutions du système $\boldsymbol{u}A \geqslant \boldsymbol{0}$ et considérons le système d'inéquations

$$\boldsymbol{u}_1\boldsymbol{b} \geqslant 0, \ldots, \boldsymbol{u}_N\boldsymbol{b} \geqslant 0 \tag{9}$$

par rapport à $\boldsymbol{b}$. Le vecteur $\boldsymbol{b}$ vérifie ce système si et seulement s'il appartient à $\mathscr{M}$.

En effet, si $\boldsymbol{ub} \geqslant 0$ pour toutes les solutions du système $\boldsymbol{u}A \geqslant \boldsymbol{0}$, le système (9) est en particulier aussi vérifié. Inversement, une solution arbitraire $\boldsymbol{u}$ du système $\boldsymbol{u}A \geqslant \boldsymbol{0}$ peut être représentée sous la forme $\boldsymbol{u} = \alpha_1\boldsymbol{u}_1 + \ldots + \alpha_N\boldsymbol{u}_N$, où tous les $\alpha_1, \ldots, \alpha_N$ sont positifs. Donc, (9) entraîne l'inéquation $\boldsymbol{ub} \geqslant 0$ pour un $\boldsymbol{u}$ arbitraire.

Ainsi, le système d'inéquations (9) définit l'ensemble $\mathscr{M}$, et le théorème est démontré.

Remarquons que dans la démonstration on aurait pu choisir au lieu de $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N$ une famille quelconque de matrices-colonnes qui engendre le cône des solutions du système $\boldsymbol{u}A \geqslant \boldsymbol{0}$. En prenant la famille fondamentale de solutions, on a obtenu pour $\mathscr{M}$ un système contenant un nombre minimal d'inéquations.

Pour un système d'équations linéaires homogènes il est facile de dégager un système équivalent composé d'un nombre minimal d'équations : il suffit pour cela de prendre les équations correspondant aux lignes du mineur principal de la matrice du système. Pour un système d'inéquations, il est plus difficile de le réaliser. L'affaire se réduit à la recherche de la carcasse d'un cône auxiliaire $\mathscr{K}^*$ engendré par les matrices-lignes des coefficients d'inéquations du système de départ. On procédera dans la suite à l'étude détaillée de ces cônes.

4. Cônes duals. Soit $\mathscr{K}$ un cône dans l'espace vectoriel $\mathscr{L}_n$. Considérons dans l'espace dual $\mathscr{L}_n^*$ de l'espace $\mathscr{L}_n$ l'ensemble $\mathscr{K}^*$ composé de tous les y pour lesquels *)

$$\langle y, x \rangle \geqslant 0, \tag{10}$$

quel que soit le vecteur $x \in \mathscr{K}$. Si les vecteurs $x_1, \ldots, x_N$ définissent la carcasse du cône $\mathscr{K}$, la condition (10) est évidemment équivalente au système d'inéquations linéaires

$$\langle y, x_i \rangle \geqslant 0, \quad i = 1, \ldots, N.$$

Ainsi, $\mathscr{K}^*$ est un cône.

DÉFINITION. Le cône $\mathscr{K}^*$ dans l'espace $\mathscr{L}_n^*$ défini par la condition (10) est dit *dual* du cône $\mathscr{K}$ dans $\mathscr{L}_n$.

Supposons que le cône $\mathscr{K}$ est défini dans une base par le système d'inéquations linéaires (2). Alors les lignes de la matrice A sont les lignes de coordonnées des vecteurs $a^1, \ldots, a^m$ de $\mathscr{L}_n^*$. Il va de soi que tous ces vecteurs appartiennent à $\mathscr{K}^*$.

Si $y \in \mathscr{K}^*$, c'est-à-dire si $\langle y, x \rangle \geqslant 0$ pour tous les $x \in \mathscr{K}$, il existe selon le théorème de Farkas des coefficients positifs $u_1, \ldots, u_m$ avec lesquels y se décompose suivant $a^1, \ldots, a^m$. Il en decoule la

PROPOSITION 16. *Les fonctions linéaires se trouvant dans les premiers membres des inéquations définissant le cône $\mathscr{K}$ engendrent le cône dual $\mathscr{K}^*$.*

Considérons le cône $\mathscr{K}^{**}$ dans $\mathscr{L}_n$, qui est dual du cône $\mathscr{K}^*$. De par la définition, $\mathscr{K}^{**}$ est l'ensemble de tous les vecteurs $x \in \mathscr{L}_n$ tels que $\langle y, x \rangle \geqslant 0$ pour tous les $y \in \mathscr{K}^*$. Il découle immédiatement de la définition que $\mathscr{K} \subseteq \mathscr{K}^{**}$. En se servant de la proposition 16, il n'est pas difficile de montrer que

$$\mathscr{K}^{**} = \mathscr{K}. \tag{11}$$

En effet, si $x_1, \ldots, x_N$ engendrent le cône $\mathscr{K}$, alors $\mathscr{K}^*$ est défini par le système d'inéquations linéaires $\langle y, x_i \rangle \geqslant 0$, $i = 1, \ldots, N$, et par suite, $\mathscr{K}^{**}$ est engendré par les mêmes vecteurs $x_1, \ldots, x_N$.

La formule (11), ainsi que la proposition (16), est un corollaire immédiat du théorème de Farkas. Inversement, ce théorème peut être obtenu facilement de la formule (11). Aussi, considère-t-on parfois la formule (11) comme l'énoncé du théorème de Farkas.

Posons que les vecteurs $x_1, \ldots, x_N$ de $\mathscr{L}_n$ engendrent le cône $\mathscr{K}$ et que les

---

*) On suppose que y représente une matrice-ligne et x une matrice-colonne, de sorte que $\langle y, x \rangle = yx$.

vecteurs $y_1, \ldots, y_M$ de $\mathscr{L}_n^*$ engendrent $\mathscr{K}^*$. Considérons les nombres

$$\nu_{ij} = \langle y_j, x_i \rangle, \quad i = 1, \ldots, N, \quad j = 1, \ldots, M. \tag{12}$$

Ils peuvent servir à composer une matrice H appelée *matrice de définition double* du cône $\mathscr{K}$. Les matrices-colonnes de H correspondent aux inéquations du système définissant $\mathscr{K}$, et les matrices-lignes aux vecteurs engendrant $\mathscr{K}$. Le cône $\mathscr{K}^*$ possède une matrice de définition double ${}^tH$.

Il ne faut pas oublier que pour le cône considéré il existe plusieurs matrices de définition double, y compris des matrices de dimensions différentes.

Si l'un des vecteurs $y_1, \ldots, y_M$ s'exprime linéairement au moyen des autres, la même dépendance est observée pour les colonnes de la matrice H. L'assertion inverse est vraie à condition que parmi $x_1, \ldots, x_N$ il y ait n vecteurs linéairement indépendants, c'est-à-dire que le cône $\mathscr{K}$ soit de dimension n. Des assertions analogues sont aussi vraies pour les lignes de la matrice H.

Dans la matrice de définition double, on s'intéresse habituellement aux éléments nuls. Pour décrire le cône, la grandeur des éléments strictement positifs n'est pas aussi essentielle.

Les colonnes nulles de la matrice H correspondent aux contraintes-égalités du système définissant le cône. En effet, si une inéquation est vérifiée en tant qu'égalité pour tous les $x_1, \ldots, x_N$, il en est de même pour toute leur combinaison linéaire.

Il est aussi évident que les lignes nulles correspondent aux x_i qui appartiennent à la face minimale du cône $\mathscr{K}$.

Ainsi, la matrice de définition double d'un cône pointé de dimension n ne renferme pas de lignes et colonnes nulles. De ce qu'on vient de dire il découle la

PROPOSITION 17. *Si le cône $\mathscr{K}$ est pointé, le cône $\mathscr{K}^*$ est de dimension n. Si le cône $\mathscr{K}$ est de dimension n, le cône $\mathscr{K}^*$ est pointé.*

Admettons maintenant que le cône $\mathscr{K}$ est pointé et de dimension n (par suite, il en est de même du cône $\mathscr{K}^*$). Dans ce cas, les carcasses de $\mathscr{K}$ et $\mathscr{K}^*$ sont constituées d'arêtes. Composons la matrice de définition double en nous servant des vecteurs directeurs des arêtes.

Chaque arête de $\mathscr{K}$ vérifie $n - 1$ inéquations linéairement indépendantes en tant qu'égalités et au moins une inéquation en tant qu'inégalité stricte. Aussi, dans la matrice de définition double H y a-t-il dans chaque ligne au moins $n - 1$ zéros situés dans les colonnes linéairement indépendantes et au moins un élément strictement positif. Il en est de même des colonnes de H car ce sont les lignes de la matrice de définition double du cône $\mathscr{K}^*$, composée pour ses arêtes. On a donc la

PROPOSITION 18. *Chaque arête du cône pointé n-dimensionnel $\mathscr{K}$ est située sur une droite qui est l'intersection de n − 1 sous-espaces tels que les demi-espaces qu'ils limitent contiennent $\mathscr{K}$. Chacun de ces sous-espaces passe par n − 1 arêtes du cône $\mathscr{K}$, dont les vecteurs directeurs sont linéairement indépendants.*

5. Théorème de séparation. Considérons encore un résultat intimement lié au théorème de Farkas. C'est le théorème de séparation pour les cônes polyédriques convexes.

THÉORÈME 4. *Si le vecteur x_0 n'appartient pas au cône polyédrique convexe fermé $\mathscr{K}$, il existe un sous-espace $(n - 1)$-dimensionnel $\mathscr{L}_{n-1}$ tel que x_0 se trouve dans l'un des demi-espaces défini par $\mathscr{L}_{n-1}$, tandis que le cône $\mathscr{K}$ est situé dans l'autre, et x_0 n'appartient pas à $\mathscr{L}_{n-1}$.*

Pour démontrer ce théorème, remarquons que, selon la formule (11), x_0 n'appartient pas à $\mathscr{K}^{**}$. Cela signifie qu'il existe un $y \in \mathscr{K}^*$ pour lequel $\langle y, x_0 \rangle < 0$. Si $y \in \mathscr{K}^*$, on a $\langle y, x \rangle \geqslant 0$ pour tout $x \in \mathscr{K}$. Ainsi donc, le sous-espace dont on démontre l'existence est le sous-espace composé des vecteurs x satisfaisant à la condition $\langle y, x \rangle = 0$.

Il existe plusieurs variantes de ce résultat, et l'on peut même fournir un tel exposé de la théorie des systèmes d'inéquations linéaires homogènes où les théorèmes de séparation se démontrent directement, alors le théorème de Farkas et les autres théorèmes sont leurs conséquences.

6. Construction de la solution générale. On étudiera ici un des procédés de construction de la solution générale d'un système homogène d'inéquations linéaires. En se servant d'une interprétation géométrique, on peut décrire ce procédé de la façon suivante.

Une inéquation non triviale définit un demi-espace dont la carcasse peut être construite facilement. Supposons qu'on a déjà construit le cône $\mathscr{K}_s$ des solutions du système de s inéquations et que la carcasse de ce cône est connue. Joignons à ce système encore une inéquation. Le demi-espace $\mathscr{M}$ qu'elle définit sépare du cône la partie

$$\mathscr{K}_{s+1} = \mathscr{K}_s \cap \mathscr{M},$$

qui est justement le cône des solutions du système de $s + 1$ inéquations. Ceci étant, dans le cas général, certaines arêtes du cône $\mathscr{K}_s$ ne sont plus incluses dans $\mathscr{M}$ et deviennent « séparées ». En revanche, on obtient dans $\mathscr{K}_{s+1}$ les nouvelles arêtes situées à l'intersection du sous-espace frontière $\mathscr{L}$ du demi-espace $\mathscr{M}$ avec les faces du cône $\mathscr{K}_s$. En joignant ainsi une à une toutes les inéquations, on obtient le système fondamental de vecteurs du cône qu'il nous faut. On étudiera plus loin en détail ce processus.

Considérons le système de s inéquations linéaires homogènes

$$a^k x = a_1^k x^1 + \dots + a_n^k x^n \geqslant 0, \quad k = 1, \dots, s.$$

définissant le cône $\mathcal{K}_s$, et soit $x_1, \ldots, x_N$ une famille de génératrices (non nécessairement minimale) de ce cône. Cherchons l'intersection du cône $\mathcal{K}_s$ avec le demi-espace $\mathcal{M}$ défini par l'inéquation

$$bx = b_1x^1 + \ldots + b_nx^n \geqslant 0.$$

Notons β_j les nombres bx_j, $j = 1, \ldots, N$, et rapportons le numéro j à l'une des trois classes I_+, I_- ou I_0 suivant que le nombre β_j est strictement positif, négatif ou nul.

Si l'ensemble I_- est vide, $bx_j \geqslant 0$ pour tous les j et on constate facilement que $\mathcal{K}_s \subseteq \mathcal{M}$.

Supposons que c'est l'ensemble I_+ qui est vide. Considérons une combinaison linéaire des génératrices à coefficients positifs $x = \alpha_1x_1 + \ldots + \alpha_Nx_N$. Si $\alpha_j > 0$ pour au moins un j de I_-, on a $bx < 0$. Il en découle que $\mathcal{K}_{s+1}$ coïncide avec l'ensemble des combinaisons linéaires positives des vecteurs x_j pour tous les $j \in I_0$.

Ainsi, on peut considérer que les deux ensembles I_+ et I_- ne sont pas vides. Soient $i \in I_+$ et $j \in I_-$. Considérons le vecteur

$$x_{ij} = \beta_jx_j - \beta_jx_i.$$

On voit aussitôt qu'il vérifie l'égalité $bx_{ij} = 0$ et, par suite, appartient au sous-espace frontière $\mathcal{L}$ du demi-espace $\mathcal{M}$. Tout vecteur $x \in \mathcal{K}_s \cap \mathcal{M}$ se décompose en une combinaison linéaire positive

$$x = \alpha_1x_1 + \ldots + \alpha_Nx_N.$$

Un couple de termes $\alpha_ix_i + \alpha_jx_j$ de cette combinaison peut être représenté en fonction de x_{ij} sous la forme

$$\alpha_ix_i + \alpha_jx_j = -\frac{\alpha_i}{\beta_j}x_{ij} + \frac{1}{\beta_j}(\alpha_i\beta_i + \alpha_j\beta_j)x_j$$

ou, si l'on exprime x_j en fonction de x_i et x_{ij}, sous la forme

$$\alpha_ix_i + \alpha_jx_j = \frac{\alpha_j}{\beta_i}x_{ij} + \frac{1}{\beta_i}(\alpha_i\beta_i + \alpha_j\beta_j)x_i.$$

Vu que $\beta_i > 0$ et $\beta_j < 0$, les coefficients dans le second membre de l'une de ces expressions sont positifs pour tous α_i et α_j positifs.

En remplaçant le couple de termes $\alpha_ix_i + \alpha_jx_j$ par sa décomposition positive en x_i et x_{ij} ou en x_j et x_{ij}, on obtient la décomposition suivante du vecteur x :

$$x = \sum_{k \in I'_+} \alpha_kx_k + \sum_{l \in I'_-} \alpha_lx_l + \sum_{h \in I_0} \alpha_hx_h + \gamma x_{ij}.$$

Ici le nombre total d'indices dans les ensembles I'_+ et I'_- est d'une unité inférieur à celui des ensembles I_+ et I_-.

Si les deux ensembles I'_+ et I'_- ne sont pas vides, on peut appliquer la même procédure à la décomposition obtenue. En substituant certains couples de termes, on aboutit à la décomposition de la forme

$$x = \sum_{k \in J_+} \alpha_k x_k + \sum_{l \in J_-} \alpha_l x_l + \sum_{h \in I_0} \alpha_h x_h + \sum_{i,\, j \in P} \gamma_{ij} x_{ij},$$

où J_+ et J_- sont les ensembles de numéros des vecteurs restants, avec $\beta_k > 0$ et $\beta_l < 0$ respectivement ; P est l'ensemble des couples de numéros étudiés. Le processus peut continuer tant que l'un des ensembles J_+ ou J_- ne devienne vide.

Si le vecteur x appartient au demi-espace $\mathscr{M}$ et que J_- ne soit pas vide, l'ensemble J_+ ne peut être vide. En effet, si J_+ était vide et J_- non vide, on aurait

$$bx = \sum_{l \in J_-} \alpha_l \beta_l < 0$$

car tous les x_{ij} et x_h pour $h \in I_0$ se trouvent dans le sous-espace $\mathscr{L}$.

Ainsi, dans tous les cas, l'ensemble J_- obtenu à la fin du processus est vide, et l'on a démontré que le vecteur $x \in \mathscr{K}_s \cap \mathscr{M}$ peut être représenté comme une combinaison linéaire positive de vecteurs x_i pour $i \in I_+ \cup I_0$ et de vecteurs x_{ij} pour $i \in I_+, j \in I_-$.

D'autre part, il est évident que toutes ces combinaisons linéaires appartiennent à l'intersection $\mathscr{K}_s \cap \mathscr{M}$. On a ainsi obtenu la

Proposition 19. *Si* $\{x_1, \ldots, x_N\}$ *est une famille de génératrices du cône* $\mathscr{K}_s$, *alors* x_i, $i \in I_+ \cup I_0$, *et* x_{ij}, $i \in I_+, j \in I_-$, *forment une famille de génératrices du cône* $\mathscr{K}_{s+1}$.

Cette proposition peut être utilisée à la recherche de la solution générale du système d'inéquations linéaires homogènes si l'on joint successivement des inéquations à la première inéquation du système. Pour commencer, on peut se servir de l'inégalité triviale identiquement vérifiée $0x \geqslant \mathbf{0}$, mais il n'est pas difficile d'indiquer une famille fondamentale de solutions pour une inéquation non triviale. On a montré dans la proposition 11 comment construire la carcasse d'un sous-espace. Démontrons maintenant la proposition suivante.

Proposition 20. *La carcasse du demi-espace défini par l'inéquation*

$$ax = a_1 x^1 + \ldots + a_n x^n \geqslant 0$$

s'obtient par adjonction du vecteur $'a = '\| a_1, \ldots, a_n \|$ *à la carcasse du sous-espace frontière* $\mathscr{L}$.

Il est évident que $'a$ appartient au demi-espace et n'appartient pas au sous-espace frontière. Considérons un vecteur arbitraire x pour lequel $ax = \alpha \geqslant 0$. Ce vecteur peut être représenté sous la forme

$$x = \frac{\alpha 'a}{a'a} + y,$$

où $y \in \mathscr{L}$. En effet, dans ce cas $ay = 0$, vu que

$$ax = \alpha = \frac{\alpha a'a}{a'a} + ay.$$

Donc, $'a$ et la carcasse de $\mathscr{L}$ constituent la famille complète des solutions de l'inéquation $ax \geqslant 0$. Il n'est pas difficile de montrer que cette famille cesse d'être complète sans l'un quelconque de ses vecteurs.

Il faut remarquer que l'application directe de la proposition, lorsqu'on joint à la famille complète donnée tous les x_{ij}, $i \in I_+$ et $j \in I_-$, conduit à des familles complètes contenant un trop grand nombre de vecteurs. Décrivons un procédé permettant de diminuer le nombre de vecteurs introduits à chaque étape. Pour simplifier, limitons-nous au cas d'un cône pointé de dimension n. On trouvera l'exposé du cas général dans le livre de Tchernikov [36].

Considérons l'intersection $\mathscr{K}_{s+1}$ d'un cône pointé n-dimensionnel $\mathscr{K}_s$ avec un demi-espace $\mathscr{M}$ de sous-espace frontière $\mathscr{L}$. Le système d'inéquations de $\mathscr{K}_{s+1}$ est obtenu par adjonction de l'inéquation

$$a_1^{s+1}x^1 + \ldots + a_n^{s+1}x^n \geqslant 0$$

définissant $\mathscr{M}$ au système d'inéquations du cône $\mathscr{K}_s$. Ainsi, le système d'inéquations

$$\begin{gathered} a_1^1x^1 + \ldots + a_n^1x^n \geqslant 0, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_1^{s+1}x^1 + \ldots + a_n^{s+1}x^n \geqslant 0 \end{gathered} \tag{13}$$

définit $\mathscr{K}_{s+1}$. Toutes les arêtes de $\mathscr{K}_{s+1}$ appartiennent à l'ensemble des demi-droites qui annulent $n - 1$ inéquations linéairement indépendantes du système (13). Ces demi-droites sont les arêtes de $\mathscr{K}_s$ et les intersections de $\mathscr{L}$ avec les faces bidimensionnelles de $\mathscr{K}_s$ ne se trouvant pas entièrement dans $\mathscr{L}$. Donc, la carcasse de $\mathscr{K}_{s+1}$ se déduit de celle de $\mathscr{K}_s$ par élimination des arêtes ne se trouvant pas dans $\mathscr{M}$ et par adjonction des intersections de $\mathscr{L}$ avec les faces bidimensionnelles de $\mathscr{K}_s$ qui ne sont pas contenues dans $\mathscr{L}$.

Il se pose le problème de déterminer lesquelles des arêtes du cône $\mathscr{K}_s$ sont voisines au sens qu'il existe une face bidimensionnelle qui passe par elles. On peut le faire en recourant à la matrice H de définition double du cône $\mathscr{K}_s$ si les lignes de la matrice correspondent aux arêtes. Une face bidi-

mensionnelles se définit par conversion en égalités exactes des $n - 2$ inéquations linéairement indépendantes du système. Aussi, les lignes de la matrice H correspondant aux arêtes voisines possèdent-elles des zéros dans les mêmes $n - 2$ colonnes, ces colonnes étant linéairement indépendantes.

Il n'est pas difficile de vérifier que dans ce cas les éléments de ces $n - 2$ colonnes ne sont simultanément nuls dans aucune autre ligne. En effet, cela signifierait que trois arêtes différentes du cône sont incluses dans une même face bidimensionnelle, ce qui contredit la proposition 12.

Notons qu'après avoir trouvé les arêtes du cône $\mathcal{K}_{s+1}$ on peut construire sa matrice de définition double et continuer à joindre les inéquations.

L'hypothèse que le cône $\mathcal{K}_s$ est pointé et de dimension n ne sera restrictive qu'aux premières étapes du processus, tant que le nombre total d'inéquations linéairements indépendantes est strictement inférieur à la dimension de l'espace. Cette hypothèse peut aussi s'avérer restrictive à toute étape s'il apparaît des contraintes-égalités.

Le processus peut débuter avec le cône pointé si le système donné d'inéquations linéaires contient un sous-système de n inéquations indépendantes dont la famille fondamentale de solutions peut facilement être obtenue ; en particulier, si le système contient les inégalités $x^1 \geqslant 0, \ldots, x^n \geqslant 0$.

Sous ce rapport, il faut remarquer qu'en augmentant le nombre de variables, on peut transformer le système d'inéquations linéaires homogènes en un système à variables positives. Pour cela, au lieu de chaque variable x^i non soumise à la condition $x^i \geqslant 0$ il faut introduire deux variables positives y^i et z^i et porter $y^i - z^i$ dans le système au lieu de x^i. D'après la solution générale du système transformé on peut facilement construire la solution générale du système initial.

§ 2. Systèmes d'inéquations linéaires non homogènes

1. Ensembles convexes dans un espace affine. On étudiera dans ce paragraphe les systèmes d'inéquations linéaires non homogènes de la forme

$$Ax \geqslant b, \tag{1}$$

ou sous une forme plus détaillée

$$\begin{gathered} a_{11}x^1 + \ldots + a_{1n}x^n \geqslant b^1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x^1 + \ldots + a_{mn}x^n \geqslant b^m. \end{gathered}$$

Comme dans le cas de systèmes homogènes on ne restreint pas la généralité en supposant que toutes les inéquations sont écrites avec le signe $\geqslant$.

En recherchant une interprétation géométrique convenable pour les solutions des systèmes d'inéquations non homogènes, on peut examiner un

cas particulier bien connu, celui des systèmes d'équations linéaires. Les systèmes d'équations linéaires non homogènes s'interprètent le plus naturellement dans un espace affine (ponctuel) (voir § 1, ch. IX) rapporté à un repère cartésien $\{O, \boldsymbol{e}\}$, dans lequel à chaque point X est associé son rayon vecteur $\overline{OX}$ et sa colonne de coordonnées $\boldsymbol{x}$ dans $\boldsymbol{e}$, *dite aussi colonne de* coordonnées du point X.

Les solutions du système d'équations linéaires non homogènes se représentent dans ce cas par des points, et l'ensemble de ces solutions par un plan. Il est naturel de représenter les solutions du système homogène associé par des vecteurs. Dans ce cas, l'ensemble de toutes les solutions du système homogène associé est le sous-espace directeur du plan défini par le système non homogène.

On agira de même avec les systèmes d'inéquations linéaires non homogènes en représentant leurs solutions par des points de l'espace affine.

Ce point est consacré à une classe importante d'ensembles dans l'espace affine, les ensembles convexes.

Admettons que le système d'inéquations linéaires (1) est compatible et que $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_s$ sont ses solutions. Si les coefficients $\alpha_1, \ldots, \alpha_s$ satisfont aux conditions

$$\alpha_1 \geqslant 0, \ldots, \alpha_s \geqslant 0, \quad \alpha_1 + \ldots + \alpha_s = 1, \tag{2}$$

la combinaison linéaire

$$\boldsymbol{x} = \alpha_1 \boldsymbol{x}_1 + \ldots + \alpha_s \boldsymbol{x}_s \tag{3}$$

est également une solution du système (1). Ces combinaisons linéaires de colonnes dont les coefficients satisfont à la condition (2) sont appelées *combinaisons convexes*. Le point X dont la colonne de coordonnées est une combinaison convexe des colonnes de coordonnées des points $X_1, \ldots, X_s$ est appelé *combinaison convexe de ces points*.

Notons que tout cela est vrai, quel que soit le repère choisi. Il va de soi que les coefficients de la combinaison linéaire ne varient pas dans tout changement de base et que de plus l'origine du repère peut être déplacée d'un vecteur quelconque. En effet, il résulte de (3) que

$$\boldsymbol{x} + \boldsymbol{p} = \alpha_1(\boldsymbol{x}_1 + \boldsymbol{p}) + \ldots + \alpha_s(\boldsymbol{x}_s + \boldsymbol{p}).$$

Soit $\mathscr{P}$ un ensemble de points de l'espace affine. On appelle *enveloppe convexe* de l'ensemble $\mathscr{P}$ l'ensemble de toutes les combinaisons convexes des points de $\mathscr{P}$.

EXEMPLES. 1) Montrons que le segment X_1X_2 est l'enveloppe convexe de ses extrémités. Le segment se trouve sur la droite d'équation paramétrique

$$\boldsymbol{x} = \boldsymbol{x}_1 + t(\boldsymbol{x}_2 - \boldsymbol{x}_1),$$

où $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ sont les colonnes de coordonnées des extrémités du segment.

Ceci étant, un point de paramètre t appartient au segment si et seulement si $t \in [0, 1]$. Aussi, en transformant l'équation de la droite à la forme

$$x = \alpha_1 x_1 + \alpha_2 x_2,$$

obtient-on $\alpha_1 = 1 - t \geqslant 0$, $\alpha_2 = t \geqslant 0$ et $\alpha_1 + \alpha_2 = 1$, ce qu'il fallait montrer.

2) Supposons que les points X_1, X_2 et X_3 ne se trouvent pas sur une même droite. Choisissons un quatrième point O non situé dans le plan bidimensionnel $\mathscr{S}_2$ passant par X_1, X_2 et X_3. Dans le plan tridimensionnel engendré par les points O, X_1, X_2, X_3, choisissons un repère d'origine O et de base $\{\overrightarrow{OX}_1, \overrightarrow{OX}_2, \overrightarrow{OX}_3\}$. Dans ce repère, le plan $\mathscr{S}_2$ a pour équation $\alpha_1 + \alpha_2 + \alpha_3 = 1$. Aussi l'intersection du plan $\mathscr{S}_2$ avec l'octant positif du repère considéré est-elle justement l'enveloppe convexe des points X_1, X_2, X_3. Ainsi, l'enveloppe convexe de trois points non alignés est un triangle de sommets en ces points.

3) Supposons que les points X_1, X_2, X_3 sont alignés et que X_2 se trouve entre X_1 et X_3. Montrons que leur enveloppe convexe est le segment X_1X_3. En effet, suivant l'hypothèse faite, il existe un nombre $\alpha \in [0, 1]$ tel que $x_2 = \alpha x_1 + (1 - \alpha)x_2$. Soit $y = \lambda x_1 + \mu x_2 + \nu x_3$, où λ, μ et ν sont positifs et leur somme vaut 1. Alors

$$y = (\lambda + \alpha\mu)x_1 + (\nu + (1 - \alpha)\mu)x_2.$$

Il est évident que les coefficients de cette combinaison sont positifs et

$$\lambda + \alpha\mu + \nu + (1 - \alpha)\mu = 1,$$

ce qui démontre la proposition.

Un ensemble de points de l'espace affine est dit *convexe* si avec tout couple de ses points il contient le segment dont ces points sont les extrémités. Remarquons que l'ensemble vide et le singleton sont des ensembles convexes.

Il est aisé de constater que sont convexes : l'espace entier, le demi-espace limité par un hyperplan arbitraire (voir § 1, ch. IX), le plan de dimension quelconque, en particulier, la droite. Le disque dans le plan euclidien est un ensemble convexe, mais le cercle ne l'est pas.

Il découle directement de la définition que l'intersection de toute famille (non nécessairement finie) d'ensembles convexes est convexe.

PROPOSITION 1. *L'enveloppe convexe de tout ensemble est un ensemble convexe.*

En effet, supposons que les points Z_1 et Z_2 appartiennent à l'enveloppe convexe de l'ensemble $\mathscr{P}$. Cela signifie que leurs colonnes de coordonnées peuvent être représentées sous forme de combinaisons convexes

$$z_1 = \alpha_1 x_1 + \ldots + \alpha_s x_s$$

et
$$z_2 = \beta_1 y_1 + \ldots + \beta_t y_t$$
de colonnes de coordonnées de points appartenant à $\mathscr{P}$. Si les nombres λ et μ sont tels que $\lambda \geqslant 0$, $\mu \geqslant 0$ et $\lambda + \mu = 1$, on a
$$\lambda z_1 + \mu z_2 = \sum_{i=1}^{s} (\lambda\alpha_i) x_i + \sum_{j=1}^{t} (\mu\beta_j) y_j .$$
Ceci étant, tous les coefficients sont ici positifs et
$$\sum_{i=1}^{s} (\lambda\alpha_i) + \sum_{j=1}^{t} (\mu\beta_j) = \lambda + \mu = 1.$$
Ainsi, le segment d'extrémités Z_1 et Z_2 est de même contenu dans l'enveloppe convexe de l'ensemble $\mathscr{P}$, si bien que la proposition est démontrée.

PROPOSITION 2. *Si les points $X_1, \ldots, X_s$ appartiennent à l'ensemble convexe $\mathscr{Q}$, toute leur combinaison convexe appartient aussi à cet ensemble.*

Démontrons cette proposition par récurrence sur le nombre de points. Pour deux points l'assertion se confond avec la définition de l'ensemble convexe (voir exemple 1)).

Soit une combinaison convexe X des points $X_1, \ldots, X_k$. Sa colonne de coordonnées x s'exprime en fonction des colonnes de coordonnées de ces points par la combinaison convexe
$$x = \alpha_1 x_1 + \ldots + \alpha_k x_k .$$
Si $\alpha_k = 0$, l'assertion se réduit immédiatement à l'hypothèse de récurrence. Si $\alpha_k = 1$, les coefficients sont nuls, si bien que l'assertion est triviale.

Dans le cas général de $0 < \alpha_k < 1$, introduisons les nombres
$$\beta_j = \frac{\alpha_j}{1 - \alpha_k}, \quad j = 1, \ldots, k - 1.$$
Ils sont tous positifs et de plus
$$\sum_{j=1}^{k-1} \beta_j = \frac{1}{1 - \alpha_k} \sum_{j=1}^{k-1} \alpha_j = \frac{1}{1 - \alpha_k} (1 - \alpha_k) = 1.$$
Par conséquent, selon l'hypothèse de récurrence, $y = \beta_1 x_1 + \ldots$ $\ldots + \beta_{k-1} x_{k-1} \in \mathscr{Q}$. En vertu de la convexité de $\mathscr{Q}$ on a alors $\alpha_k x_k +$ $+ (1 - \alpha_k) y \in \mathscr{Q}$, c'est-à-dire que
$$\alpha_1 x_1 + \ldots + \alpha_{k-1} x_{k-1} + \alpha_k x_k \in \mathscr{Q},$$
ce qu'il fallait démontrer.

Il découle de la proposition 2 que tout ensemble convexe $\mathscr{Q}$ qui contient l'ensemble $\mathscr{P}$ contient aussi son enveloppe convexe. Ce fait et la proposition 1 entraînent la

PROPOSITION 3. *L'enveloppe convexe de l'ensemble $\mathscr{P}$ est l'intersection de tous les ensembles convexes contenant $\mathscr{P}$.*

Un ensemble $\mathscr{P}$ de points de l'espace affine est dit *borné* si dans un repère toutes les coordonnées de tous ses points sont en valeur absolue bornées par un même nombre. Ceci est équivalent au fait que l'ensemble des colonnes de coordonnées des points de $\mathscr{P}$ est borné en c-norme. En vertu du théorème 1, § 3, ch. XI, on peut en conclure qu'un ensemble dans l'espace affine est borné si et seulement si l'ensemble correspondant des colonnes de coordonnées est borné en une norme quelconque.

Il n'est pas difficile de montrer que la définition donnée est indépendante du choix du repère. En effet, quand on change de repère, on remplace la colonne de coordonnées $\boldsymbol{x}$ par $\boldsymbol{x}' = S\boldsymbol{x} + \boldsymbol{p}$, et l'on peut écrire l'estimation

$$\|\boldsymbol{x}'\|_\infty \leqslant \|S\|_c \|\boldsymbol{x}\|_\infty + \|\boldsymbol{p}\|_\infty$$

qui entraîne l'assertion nécessaire.

PROPOSITION 4. *Si l'ensemble $\mathscr{P}$ est borné, il en est de même de son enveloppe convexe. En particulier, l'enveloppe convexe d'un ensemble fini est bornée.*

DÉMONSTRATION. Pour la colonne de coordonnées d'un point quelconque de l'enveloppe convexe on a la décomposition

$$\boldsymbol{y} = \alpha_1 \boldsymbol{x}_1 + \ldots + \alpha_s \boldsymbol{x}_s$$

qui est une combinaison convexe des colonnes de coordonnées des points de $\mathscr{P}$. D'où il vient

$$\|\boldsymbol{y}\| \leqslant \sum_{i=1}^{s} \alpha_i \|\boldsymbol{x}_i\| \leqslant \max_i \|\boldsymbol{x}_i\| \sum_{i=1}^{s} \alpha_s = \max_i \|\boldsymbol{x}_i\|.$$

Etant donné que $\mathscr{P}$ est borné, il existe un nombre ρ tel que $\|\boldsymbol{x}\| \leqslant \rho$ pour tous les $X \in \mathscr{P}$. Donc, $\max_i \|\boldsymbol{x}_i\| \leqslant \rho$ et $\|\boldsymbol{y}\| \leqslant \rho$, ce qu'il fallait démontrer.

Rapportons l'espace affine à un repère choisi. Toute inéquation linéaire à coefficients non nuls définit alors un demi-espace fermé par rapport au plan ne passant en général pas par l'origine du repère. Le système de ces inéquations définit l'intersection des demi-espaces correspondants. L'inéquation triviale (dont le premier membre est nul) est soit incompatible, soit vérifiée identiquement. Par conséquent, l'adjonction de ces inéquations ou bien ne modifie pas l'ensemble de solutions, ou bien le rend vide.

L'intersection de demi-espaces est un ensemble convexe, vu que chaque demi-espace est convexe.

Définition. L'intersection de demi-espaces de l'espace affine s'appelle *ensemble polyédrique convexe*. Si l'ensemble polyédrique convexe est borné, on l'appelle *polyèdre convexe*.

En vertu de cette définition, l'ensemble des solutions d'un système d'inéquations linéaires non homogènes est représenté par un ensemble polyédrique convexe.

Les cônes polyédriques convexes définis au § 1 peuvent être aussi considérés comme des ensembles ponctuels dans l'espace affine. Ils constituent alors un cas spécial d'ensembles polyédriques convexes dont les plans frontières de tous les demi-espaces passent par un même point, l'origine des coordonnées.

2. Ensemble des solutions d'un système d'inéquations linéaires non homogènes. On a vu que l'ensemble des points représentant les solutions du système de la forme (1) est l'intersection de demi-espaces et par suite, est un ensemble convexe. On a même donné un nom à ces ensembles. Etudions-les maintenant plus en détail. A cet effet, introduisons une variable supplémentaire x^{n+1} et considérons le système

$$\begin{aligned} A\boldsymbol{x} - x^{n+1}\boldsymbol{b} &\geqslant \boldsymbol{0}, \\ x^{n+1} &\geqslant 0. \end{aligned} \tag{4}$$

Pour chaque solution $\boldsymbol{x}$ du système (1) la matrice-colonne $\boldsymbol{q} = {}^t\|{}^t\boldsymbol{x}, 1\|$ vérifie le système (4) et, inversement, si $x^{n+1} = 1$, les n premiers éléments de la solution du système (4) vérifient le système (1).

Notons $\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s\}$ la famille fondamentale de solutions du système (4) et admettons que $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_s$ sont numérotées de telle sorte que la dernière composante dans les matrices-colonnes $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_t$ soit strictement positive et dans les matrices-colonnes $\boldsymbol{p}_{t+1}, \ldots, \boldsymbol{p}_s$ soit égale à zéro. De plus, on peut même considérer que la dernière composante dans $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_t$ est égale à 1, vu que toute matrice-colonne peut être remplacée par une colonne proportionnelle, avec facteur de proportionnalité strictement positif.

Si le système (1) est compatible, $t \geqslant 1$. En effet, dans le cas contraire, pour tous $\boldsymbol{x}$ et x^{n+1} il ressort de $A\boldsymbol{x} - x^{n+1}\boldsymbol{b} \geqslant \boldsymbol{0}$ que $x^{n+1} = 0$ et, en particulier, pour la solution $\boldsymbol{x}$ du système (1) il découle de $A\boldsymbol{x} - \boldsymbol{b} \geqslant \boldsymbol{0}$ que $1 = 0$.

Toutes les solutions du système (4) et elles seules peuvent être représentées par la formule

$$\boldsymbol{q} = \alpha_1\boldsymbol{p}_1 + \ldots + \alpha_t\boldsymbol{p}_t + \beta_1\boldsymbol{p}_{t+1} + \ldots + \beta_{s-t}\boldsymbol{p}_s \tag{5}$$

à coefficients positifs α_i et β_j. Ceci étant, en vertu des hypothèses faites, la

dernière composante de la matrice-colonne q est égale à 1 si et seulement si

$$\alpha_1 + \dots + \alpha_t = 1.$$

L'égalité obtenue est nécessaire et suffisante pour que les n premières composantes de la matrice-colonne (5) représentent une solution du système (1). A cette condition, en rejetant les derniers éléments de toutes les matrices-colonnes dans la formule (5), on aboutit à la *solution générale du système d'inéquations linéaires* (1).

Du point de vue géométrique, l'adjonction de la variable supplémentaire x^{n+1} signifie le passage à l'espace $(n+1)$-dimensionnel dans lequel l'espace n-dimensionnel de départ est un hyperplan d'équation $x^{n+1} = 1$. Le système d'inéquations homogènes (4) définit un cône polyédrique convexe $\overline{\mathscr{K}}$. L'ensemble des solutions du système (1) est l'intersection du cône $\overline{\mathscr{K}}$ et de l'hyperplan $x^{n+1} = 1$.

Pour $x^{n+1} = 0$, le système (4) se transforme en un système homogène

$$Ax \geqslant \mathbf{0}. \tag{6}$$

Les solutions de ce système seront représentées sous forme de vecteurs. Vu que la dernière composante de chacun de ces vecteurs est égale à zéro, ces derniers se trouvent dans le sous-espace directeur de l'hyperplan $x^{n+1} = 1$. Le système (6) définit dans ce sous-espace un cône polyédrique convexe $\mathscr{K}$. On voit aussitôt que $p_{t+1}, \dots, p_s$ sont les colonnes de coordonnées des vecteurs de la carcasse de $\mathscr{K}$.

Les matrices-colonnes $p_1, \dots, p_t$ seront assimilées aux colonnes de coordonnées des points $P_1, \dots, P_t$. Alors le premier groupe de termes dans la formule (5) constitue une combinaison convexe des points $P_1, \dots, P_t$, et le second groupe, un vecteur du cône $\mathscr{K}$. Ainsi, on est en mesure de formuler le théorème suivant qui est la traduction géométrique de la formule (5).

THÉORÈME 1. *Soit $\mathscr{M}$ un ensemble polyédrique convexe. Il existe une enveloppe convexe d'une famille finie de points $\mathscr{P}$ et un cône polyédrique convexe $\mathscr{K}$ tels que tous les points de $\mathscr{M}$ et eux seuls sont les extrémités des vecteurs de $\mathscr{K}$ issus des points de $\mathscr{P}$.*

Supposons que le système (6) n'a qu'une solution triviale. Le cône $\mathscr{K}$ est alors composé du vecteur nul, de sorte que l'ensemble $\mathscr{M}$ se confond avec $\mathscr{P}$ et, par suite, est borné. Par contre, si $\mathscr{K}$ contient un vecteur non nul, en le multipliant par un coefficient suffisamment grand on peut obtenir un point de l'ensemble $\mathscr{M}$ dont certaines coordonnées sont en valeur absolue aussi grandes que l'on veut. On a donc le corollaire suivant.

COROLLAIRE. *Un ensemble polyédrique convexe défini par le système* (1) *est un polyèdre si et seulement si le système homogène associé ne possède qu'une solution triviale.*

Un autre cas particulier limite se présente pour $t = 1$. Dans ce cas l'ensemble $\mathscr{P}$ est composé d'un seul point et $\mathscr{M}$ est alors un cône polyédrique convexe, considéré comme ensemble ponctuel.

Démontrons le théorème inverse du théorème 1.

THÉORÈME 2. *Supposons que l'ensemble $\mathscr{P}$ est l'enveloppe linéaire d'une famille finie de points, $\mathscr{K}$ un cône polyédrique convexe et $\mathscr{M}$ un ensemble de points qui sont des extrémités des vecteurs de $\mathscr{K}$ issus des points de $\mathscr{P}$. $\mathscr{M}$ est alors un ensemble polyédrique convexe.*

DÉMONSTRATION. La colonne de coordonnées d'un point arbitraire X de $\mathscr{M}$ est de la forme

$$x = \alpha_1 x_1 + \ldots + \alpha_t x_t + \beta_1 l_1 + \ldots + \beta_h l_h,$$

où $x_1, \ldots, x_t$ et $l_1, \ldots, l_h$ sont les colonnes de coordonnées respectivement des points qui engendrent $\mathscr{P}$ et des vecteurs qui engendrent $\mathscr{K}$. Tous les coefficients sont positifs et $\alpha_1 + \ldots + \alpha_t = 1$. Notons p_i, $i = 1, \ldots, t$, les matrices-colonnes déduites de x_i par adjonction de la dernière composante égale à 1, et q_j, $j = 1, \ldots, h$, les matrices-colonnes déduites de l_j par adjonction de la dernière composante égale à 0.

L'ensemble de toutes les combinaisons linéaires positives des matrices-colonnes p_i et q_j

$$\bar{x} = \alpha_1 p_1 + \ldots + \alpha_t p_t + \beta_1 q_1 + \ldots + \beta_h q_h$$

est un cône polyédrique convexe $\bar{\mathscr{K}}$ dans l'espace vectoriel de dimension $n + 1$. Selon le théorème 3 du § 1, il existe un ensemble de matrices-lignes $a^1, \ldots, a^N$ à $n + 1$ éléments, telles que $\bar{x} \in \bar{\mathscr{K}}$ si et seulement si

$$a^k \bar{x} \geqslant 0$$

pour tous les $k = 1, \ldots, N$. Ecrivons ces inéquations d'une façon plus développée et tenons compte du fait que la somme des coefficients α_i vaut l'unité. Il vient

$$a^k \bar{x} = \sum_{i=1}^{t} \alpha_i (a^k x_i) + \sum_{i=1}^{t} \alpha_i a^k_{n+1} + \sum_{j=1}^{h} \beta_j (a^k l_j) = a^k x + a^k_{n+1} \geqslant 0.$$

Or $X \in \mathscr{M}$ si et seulement si $\bar{x} \in \bar{\mathscr{K}}$. Donc, $X \in \mathscr{M}$ si et seulement si sa colonne de coordonnées vérifie le système d'inéquations linéaires non homogènes

$$a^k x + a^k_{n+1} \geqslant 0, \quad k = 1, \ldots, N.$$

Le théorème est démontré.

COROLLAIRE. *L'enveloppe convexe d'un ensemble fini de points est un polyèdre convexe.*

3. Faces d'un ensemble polyédrique convexe. On dit qu'un ensemble polyédrique convexe non vide est de *dimension d* s'il est situé dans un plan de dimension d et n'est situé dans aucun plan de dimension inférieure.

Les inéquations figurant dans le système non homogène de la forme (1) peuvent être partagées en contraintes-égalités et contraintes-inégalités, comme cela a été fait au § 1 relativement aux inéquations homogènes. Par contraintes-égalités on entend des inéquations qui se vérifient pour les solutions du système en tant qu'égalités exactes.

PROPOSITION 5. *La dimension de l'ensemble polyédrique convexe* $\mathcal{M}$ *défini par le système* (1) *est égale à* $n - \rho$, *où* ρ *est le rang de la matrice formée de tous les coefficients des contraintes-égalités.*

La démonstration s'appuie sur des raisonnements analogues à ceux utilisés lors de la démonstration des propositions 2 à 5 du § 1. En remplaçant le signe $\geqslant$ par $=$ dans les contraintes-égalités on obtient les équations du plan $\mathcal{S}_{n-\rho}$ de dimension $n - \rho$, où se trouve l'ensemble polyédrique $\mathcal{M}$. Si l'on exprime ρ variables à partir de ces équations et qu'on les porte dans les inéquations restantes, on aboutit à un système de contraintes-inégalités qui définit $\mathcal{M}$ dans le plan $\mathcal{S}_{n-\rho}$.

Considérons ensuite $\mathcal{S}_{n-\rho}$ comme un espace indépendant et montrons que l'ensemble défini par le système de contraintes-inégalités n'est situé dans aucun hyperplan. A cet effet, montrons d'abord que le système correspondant d'inéquations strictes a une solution x_0. En effet, chaque inéquation devient une inégalité stricte en un certain point de l'ensemble $\mathcal{M}$. Par conséquent, toute combinaison convexe de ces points à coefficients non nuls est la solution x_0 cherchée.

Il n'est pas difficile de démontrer que x_0 est un point intérieur à l'ensemble au sens qu'il lui appartient avec un voisinage de dimension $n - \rho$. On peut en déduire l'assertion nécessaire par analogie à la démonstration de la proposition 5 du § 1.

DÉFINITION. On appelle *face* de l'ensemble polyédrique convexe défini par le système (1) un ensemble polyédrique convexe de solutions d'un système déduit de (1) par substitution des égalités à certaines contraintes-inégalités.

Les faces unidimensionnelles sont appelées *arêtes* et les faces de dimension zéro, *sommets* de l'ensemble polyédrique convexe.

Tout ensemble polyédrique convexe unidimensionnel est soit une droite, soit une demi-droite, soit un segment. Par conséquent, si les arêtes existent, elles ne peuvent être que des trois espèces énumérées plus haut. Dans un polyèdre convexe borné, les arêtes ne peuvent être que des segments.

Un ensemble polyédrique k-dimensionnel est situé dans un plan de

dimension k et (s'il n'est pas vide et ne coïncide pas avec tout le plan) y est défini par un système d'inéquations contenant des contraintes-inégalités non triviales. Aussi l'ensemble polyédrique k-dimensionnel possède-t-il des faces $(k - 1)$-dimensionnelles obtenues par substitution d'une équation à l'une de ces contraintes. Remarquons que toutes les faces ne peuvent pas être vides. S'il en était ainsi, l'ensemble polyédrique se composerait des solutions du système des contraintes-inégalités strictes. Alors par un raisonnement anologue à celui de la démonstration de la proposition 4 du § 1 on pourrait démontrer que cet ensemble est ouvert. Or, étant une intersection d'ensembles fermés, il doit être fermé. On n'examinera pas ici ce fait en détail car il resulte directement du théorème 7 qu'on démontrera plus loin.

Ainsi donc, un ensemble polyédrique convexe non vide de dimension k, non confondu avec le plan k-dimensionnel doit admettre des faces $(k - 1)$-dimensionnelles. Mais on ne peut affirmer qu'il possède des faces de dimension inférieure, vu qu'une face $(k - 1)$-dimensionnelle peut être un plan de dimension $k - 1$. Plus précisément on a la

PROPOSITION 6. *Un ensemble polyédrique convexe défini par un système d'inéquations linéaires à matrice A de rang r ne peut avoir de faces de dimension inférieure à $n - r$.*

En effet, pour que la face soit de dimension k, elle doit transformer en équations $n - k$ inéquations linéairement indépendantes du système. Or le nombre total d'équations linéairement indépendantes déduites du système (1) est au plus égal à r.

Il découle de la proposition démontrée que pour l'existence des sommets d'un ensemble polyédrique convexe il est nécessaire que $n = r$. Démontrons que cette condition est suffisante, c'est-à-dire qu'on a la proposition suivante.

PROPOSITION 7. *Un ensemble polyédrique convexe non vide $\mathcal{M}$ défini par le système* (1) *possède des sommets si et seulement si les colonnes de la matrice du système sont linéairement indépendantes.*

DÉMONSTRATION. Soit $r = n$. Le cône $\overline{\mathcal{K}}$ défini par le système (4) est alors pointé, car le système d'équations

$$Ax - x^{n+1}b = 0,$$
$$x^{n+1} = 0$$

n'a qu'une solution triviale. Donc, chacune des matrices-colonnes p_1, ..., p_s de la formule (5) définit une arête du cône $\overline{\mathcal{K}}$ et, par suite, transforme en équations n (nombre de variables inférieur de l'unité) inéquations quelconques linéairement indépendantes du système (4). Pour les matrices-

colonnes $p_1, \ldots, p_t$ la dernière inégalité n'intervient pas parmi les n inéquations transformées, car x^{n+1} y est égal à 1. Donc, les n premiers éléments de chacune des colonnes vérifient les n inéquations linéairement indépendantes du système (1) en tant qu'égalités exactes. Il en découle que les n premiers éléments des colonnes $p_1, \ldots, p_t$ constituent les colonnes de coordonnées des sommets de l'ensemble polyédrique $\mathscr{M}$. La proposition est démontrée.

Dans l'exposé ultérieur, il ne nous faudra pas changer de repère. Aussi, pour faciliter l'exposé, identifiera-t-on un point ou vecteur à une colonne de coordonnées associée. On transforme ainsi l'espace affine en espace arithmétique dans lequel on étudiera les objets de l'espace affine : points, vecteurs, plans de différentes dimensions, etc.

De toutes les propriétés des faces de dimension arbitraire démontrons la suivante :

PROPOSITION 8. *Soit $\mathscr{M}'$ une face de l'ensemble polyédrique $\mathscr{M}$. Si un point $x \in \mathscr{M}'$ se représente comme une combinaison convexe*

$$x = \alpha_1 x_1 + \ldots + \alpha_k x_k$$

à coefficients non nuls des points $x_1, \ldots, x_k$ de $\mathscr{M}$, ces points appartiennent tous à $\mathscr{M}'$.

Pour démontrer l'assertion, considérons une inéquation arbitraire $a^i x \geqslant b^i$ du système (1), qui devient égalité sur la face $\mathscr{M}'$. On a

$$a^i x = \alpha_1 a^i x_1 + \ldots + \alpha_k a^i x_k .$$

Pour tous les j on a ici $a^i x_j \geqslant b^i$. Si pour un j_0 est vérifié $a^i x_{j_0} > b^i$, il ressort de $\alpha_1 > 0, \ldots, \alpha_k > 0$ et de $\alpha_1 + \ldots + \alpha_k = 1$ que $a^i x > b^i$. La contradiction obtenue démontre la proposition.

DÉFINITION. Un point x de l'ensemble convexe $\mathscr{M}$ sera appelé *point extrémal* si de $x = \alpha x_1 + (1 - \alpha) x_2$ pour $0 < \alpha < 1$ et $x_1, x_2 \in \mathscr{M}$ il résulte que $x = x_1 = x_2$. Autrement dit, un point extrémal n'est intérieur à aucun segment inclus dans $\mathscr{M}$.

PROPOSITION 9. *Les sommets d'un ensemble polyédrique convexe et eux seuls sont ses points extrémaux.*

DÉMONSTRATION. Si x est un sommet, c'est-à-dire une face de dimension 0 de l'ensemble $\mathscr{M}$, il découle de $x = \alpha x_1 + (1 - \alpha) x_2$, $0 < \alpha < 1$ et x_1, $x_2 \in \mathscr{M}$, en vertu de la proposition 8 que x_1 et x_2 sont situés dans la même face de dimension 0 et, par suite, coïncident avec x. Ainsi, tout sommet est un point extrémal.

Démontrons que tout point extrémal est un sommet. Vérifions d'abord que si un point extrémal existe, on a $n = \mathrm{Rg}\, A$. En effet, si y est une solu-

tion non triviale du système d'équations $Ay = \mathbf{0}$, chaque solution x du système (1) peut être considérée comme le milieu du segment d'extrémités $x + y$ et $x - y$ qui sont aussi des solutions du système (1).

Démontrons ensuite que la décomposition du point extrémal suivant la formule (5) ne contient pas de termes de numéros strictement supérieurs à t. En effet, soit x le point extrémal. Dans ce cas, la matrice-colonne $q = {}^t\| {}^t x, 1\|$ peut être représentée par la formule (5) sous la forme

$$q = \alpha_1 p_1 + \ldots + \alpha_t p_t + \beta_1 p_{t+1} + \ldots + \beta_{s-t} p_s,$$

ou $q = u + v$, avec $u = \alpha_1 p_1 + \ldots + \alpha_t p_t$ et $v = \beta_1 p_{t+1} + \ldots + \beta_{s-t} p_s$. Les matrices-colonnes $u + \frac{1}{2} v$ et $u + \frac{3}{2} v$ correspondent aux points de l'ensemble $\mathscr{M}$, et le point q est le milieu du segment limité par ces points, à condition que $v \neq \mathbf{0}$.

Ainsi, pour le point extrémal on a

$$q = \alpha_1 p_1 + \ldots + \alpha_t p_t.$$

Il en découle que q se confond avec l'un des $p_1, \ldots, p_t$. En effet, soit par exemple $\alpha_1 \neq 0$. Si $\alpha_2 + \ldots + \alpha_t = 0$, la démonstration est immédiate. Dans le cas contraire, on a $q = \alpha_1 p_1 + (1 - \alpha_1) p_0$, où

$$p_0 = (\alpha_2 + \ldots + \alpha_t)^{-1}(\alpha_2 p_2 + \ldots + \alpha_t p_t),$$

et de la définition du point extrémal il découle que $q = p_1 = p_0$.

Or on a vu en démontrant la proposition 6 que si les sommets existent, les matrices-colonnes $p_1, \ldots, p_t$ définissent justement les sommets, ce qui achève la démonstration.

Corollaire. *Le nombre des sommets d'un ensemble polyédrique convexe est fini.*

En effet, si les sommets existent, ils sont des points extrémaux. Or les points extrémaux se définissent par les matrices-colonnes $p_1, \ldots, p_t$. Donc, tous les sommets sont définis par ces colonnes.

Si l'ensemble polyédrique convexe défini par le système (1) est borné, le système homogène (6) n'a qu'une solution triviale et, par suite, $n = r$, de sorte que le polyèdre convexe possède nécessairement des sommets. On aboutit au théorème suivant.

Théorème 3. *Un polyèdre convexe possède des sommets (points extrémaux) et constitue l'enveloppe convexe de ses sommets.*

Ce théorème est un cas particulier du théorème plus général d'après lequel chaque ensemble convexe borné admet des points extrémaux et constitue l'enveloppe convexe de ces points. Dans le cas général, l'ensemble des points extrémaux est infini. Par exemple, les points extrémaux d'une boule

sont les points de la sphère frontière. On trouvera la démonstration de ce théorème général par exemple dans le livre de Nikaido [29].

Servons-nous du théorème de séparation pour les cônes polyédriques convexes et du passage d'un système d'inéquations non homogènes à un système homogène associé pour obtenir le théorème de séparation pour les ensembles polyédriques convexes.

THÉORÈME 4. *Soit* $Az < \boldsymbol{b}$. *Il existe une matrice-ligne* $\boldsymbol{v}$ *et un nombre* w *tels que* $\boldsymbol{v}z + w < 0$, *et pour toutes les solutions du système* $A\boldsymbol{x} \geqslant \boldsymbol{b}$ *on a* $\boldsymbol{v}\boldsymbol{x} + w \geqslant 0$.

DÉMONSTRATION. Notons que la matrice-colonne $\bar{z}$ déduite de z par adjonction de la dernière composante égale à 1 n'appartient pas à l'ensemble des solutions du système homogène (4) :

$$A\boldsymbol{x} - \boldsymbol{b}x^{n+1} \geqslant \boldsymbol{0},$$
$$x^{n+1} \geqslant 0.$$

Selon le théorème 4 du § 1 il existe une matrice-ligne $\bar{\boldsymbol{v}}$ à $n + 1$ éléments pour laquelle $\bar{\boldsymbol{v}}\bar{z} < 0$ et $\bar{\boldsymbol{v}}\bar{\boldsymbol{x}} \geqslant 0$ quelles que soient les solutions du système (4), en particulier celles dont la dernière composante vaut 1. Les n premières composantes de chacune de ces solutions vérifient le système $A\boldsymbol{x} \geqslant \boldsymbol{b}$. Inversement, si $\boldsymbol{x}$ est solution du système $A\boldsymbol{x} \geqslant \boldsymbol{b}$, alors $\bar{\boldsymbol{x}} = {}^t\| {}^t\boldsymbol{x}, 1 \|$ est celle du système (4).

Soit $\bar{\boldsymbol{v}} = \| \boldsymbol{v}, w \|$. Alors $\bar{\boldsymbol{v}}\bar{z} = \boldsymbol{v}z + w < 0$ et $\bar{\boldsymbol{v}}\bar{\boldsymbol{x}} = \boldsymbol{v}\boldsymbol{x} + w \geqslant 0$ pour toutes les solutions du système $A\boldsymbol{x} \geqslant \boldsymbol{b}$, ce qu'il fallait démontrer.

4. Condition de compatibilité. A la différence des systèmes homogènes les systèmes non homogènes d'inéquations linéaires peuvent être incompatibles. Il existe différentes conditions de compatibilité de ces systèmes. Les conditions généralisant le théorème de Kronecker-Capelli peuvent être trouvées dans le livre de Tchernikov [36]. On examinera ici l'une des conditions généralisant le théorème de Fredholm.

THÉORÈME 5. *Le système d'inéquations linéaires*

$$A\boldsymbol{x} \geqslant \boldsymbol{b} \tag{1}$$

est compatible si et seulement si $\boldsymbol{u}A = \boldsymbol{0}$ *et* $\boldsymbol{u} \geqslant \boldsymbol{0}$ *entraînent* $\boldsymbol{u}\boldsymbol{b} \leqslant 0$.

DÉMONSTRATION. Supposons que le système d'inéquations (1) est compatible, c'est-à-dire qu'il existe une matrice-colonne $\boldsymbol{x}$ à n éléments pour laquelle $A\boldsymbol{x} \geqslant \boldsymbol{b}$. Pour toute matrice-ligne positive $\boldsymbol{u}$ à m éléments on a alors $\boldsymbol{u}A\boldsymbol{x} \geqslant \boldsymbol{u}\boldsymbol{b}$ et, par suite, $\boldsymbol{u}A = \boldsymbol{0}$ entraîne $\boldsymbol{u}\boldsymbol{b} \leqslant 0$.

L'assertion inverse est moins évidente. Montrons que pour un système incompatible de la forme (1) il existe une matrice-ligne $\boldsymbol{u} \geqslant \boldsymbol{0}$ pour laquelle $\boldsymbol{u}A = \boldsymbol{0}$ et $\boldsymbol{u}\boldsymbol{b} > 0$. Considérons les matrices-lignes $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_s$ (famille fon-

damentale des solutions du système d'équations $uA = \mathbf{0}$) et la matrice U composée des lignes $u_1, \ldots, u_s$. Il est évident que $UA = O$.

Notons $\mathscr{S}$ le plan dans l'espace $\mathscr{R}_m$, composé de toutes les matrices-colonnes de la forme $Ax - b$, $x \in \mathscr{R}_n$. Pour tous les $y \in \mathscr{S}$ on a

$$u_i y = -u_i b \quad (i = 1, \ldots, s). \tag{7}$$

Inversement, si les équations (7) sont vérifiées pour un y, la matrice-colonne $y + b$ vérifie les hypothèses du théorème de Fredholm, si bien que le système $Ax = y + b$ est compatible. Ainsi, le système d'équations (7) définit le plan $\mathscr{S}$. Pour écrire ce système sous forme matricielle, introduisons la matrice-colonne $q = -Ub$. Alors (7) prend la forme

$$Uy = q. \tag{8}$$

Le système d'inéquations (1) est incompatible si et seulement si le plan $\mathscr{S}$ ne contient pas de matrices-colonnes positives ou, ce qui revient au même, le système (8) n'a pas de solutions positives. Dans ce cas, selon le corollaire du théorème 2, § 1, le système d'inéquations

$$zU \geqslant \mathbf{0}, \quad zq < 0$$

doit être compatible. Soit z une solution quelconque de ce système. Désignons zU par u. Alors $u \geqslant \mathbf{0}$, $uA = zUA = \mathbf{0}$ et $zq = -zUb < 0$, c'est-à-dire que $ub > 0$. Ainsi, la matrice-ligne u est celle qu'il nous fallait trouver.

Cherchons quelques corollaires du théorème démontré.

PROPOSITION 10. *La matrice A étant fixée, l'ensemble des matrices-colonnes b pour lesquelles le système* (1) *est compatible, est fermé pour toute norme sur $\mathscr{R}_m$, tandis que l'ensemble des matrices-colonnes b pour lesquelles ce système est incompatible, est ouvert.*

Il suffit de démontrer la dernière assertion car le complémentaire d'un ensemble ouvert est fermé.

Le système (1) est incompatible s'il existe une matrice-ligne positive u telle que $uA = \mathbf{0}$ et $ub > 0$. Si une matrice-colonne b' diffère peu de b, la même matrice-ligne u satisfait à la condition $ub' > 0$, d'où l'incompatibilité du système $Ax \geqslant b'$. Démontrons cette assertion en majorant en c-norme la différence $\boldsymbol{\beta} = b' - b$.

Si $ub > 0$, l'inégalité $u(b + \boldsymbol{\beta}) > 0$ est vérifiée à condition que $|u\boldsymbol{\beta}| < ub$. Mais

$$|u\boldsymbol{\beta}| \leqslant \max_i |\beta_i| \sum_{i=1}^{m} u_i = \nu \|\boldsymbol{\beta}\|_c,$$

où $\nu = \sum_{i=1}^{m} u_i$. Aussi l'écart pour lequel

$$\|\beta\|_c < \frac{ub}{\nu}$$

sera-t-il suffisamment petit. Notons que $\nu \neq 0$. En effet, les éléments de la matrice-ligne u sont positifs et parmi eux il y a des éléments non nuls car $ub > 0$.

PROPOSITION 11. *Le système d'inéquations linéaires $Ax \geqslant b$ est compatible pour tout second membre b si et seulement si le système d'équations $uA = \mathbf{0}$ n'a qu'une solution triviale positive.*

DÉMONSTRATION. Supposons que le système $uA = \mathbf{0}$ a une solution positive u et que le système $Ax \geqslant b$ est compatible pour tout second membre. Prenons pour second membre la matrice-colonne ${}^t u$. Selon le théorème 5 on doit alors avoir l'inégalité $u\,{}^t u \leqslant 0$ signifiant que $u = \mathbf{0}$.

L'assertion inverse se démontre d'une façon aussi simple.

Dans les applications, on a souvent affaire à des systèmes d'inéquations de la forme (1) où on exige que les variables x^i soient positives. Cette condition supplémentaire est évidemment équivalente à l'adjonction de n inégalités $x^i \geqslant 0$ et n'introduit en principe rien de nouveau. Cependant, il est utile de formuler la condition de compatibilité pour ce cas-là.

PROPOSITION 12. *Le système d'inéquations linéaires $Ax \geqslant b$ a une solution positive $x \geqslant \mathbf{0}$ si et seulement si $uA \leqslant \mathbf{0}$ et $u \geqslant \mathbf{0}$ entraînent $ub \leqslant 0$.*

DÉMONSTRATION. Soient $\bar{A}$ une matrice déduite de A par écriture en bas de la matrice unité d'ordre n, et $\bar{b}$ une matrice-colonne déduite de b par adjonction en bas de n éléments nuls. L'existence d'une solution positive du système (1) est équivalente à la compatibilité du système

$$\bar{A}x \geqslant \bar{b}. \tag{9}$$

Pour appliquer le théorème 5, considérons d'abord les matrices-lignes de la forme

$$\bar{u} = \| u, v \| = \| u_1, \ldots, u_m, v_1, \ldots, v_n \|.$$

Le système (9) est compatible si et seulement si $\bar{u}\bar{A} = \mathbf{0}$ et $\bar{u} \geqslant \mathbf{0}$ entraînent $\bar{u}\bar{b} \leqslant 0$. D'une façon plus développée, cela signifie que $uA + v = \mathbf{0}$ et $u \geqslant \mathbf{0}$, $v \geqslant \mathbf{0}$ entraînent $ub + v\mathbf{0} \leqslant 0$. En éliminant de l'hypothèse la ligne v qui ne figure pas dans la conclusion, on obtient l'assertion exigée.

Il existe aussi d'autres conditions de compatibilité des systèmes d'inéquations linéaires non homogènes. Comme toutes celles qui ont été mentionnées plus haut, elles conviennent parfaitement à des raisonnements

théoriques, mais ne sont guère utiles à l'étude pratique des systèmes concrets.

Les paragraphes suivants seront consacrés à l'étude d'un problème de programmation linéaire. La première étape de sa résolution est consacrée à la recherche d'un point de l'ensemble polyédrique convexe (ou à l'établissement du fait que cet ensemble est vide). Les programmes de résolution de ce problème sont accessibles, faciles à appliquer et se distinguent par un rendement efficace. Ils peuvent donc être utilisés avec succès à l'étude de la compatibilité des systèmes d'inéquations linéaires.

5. Inéquations résultant d'un système non homogène d'inéquations linéaires. Pour les systèmes non homogènes, le théorème de Farkas se présente sous la forme suivante : l'inéquation linéaire $cx \geqslant f$ est une conséquence du système compatible d'inéquations linéaires (1) si et seulement si elle est une combinaison linéaire positive d'inéquations du système, ou bien résulte de cette combinaison linéaire quand on diminue le terme constant. Dans ce dernier cas, l'inéquation est appelée *combinaison linéaire affaiblie*. On démontrera ce théorème sous une forme analogue à celle du théorème des système homogènes (théorème 2, § 1) en laissant au lecteur le soin de vérifier que cette formulation est équivalente à celle donnée plus haut.

THÉORÈME 6. *Si le système $Ax \geqslant b$ est compatible, un et un seul des deux systèmes d'inéquations*

$$Ax \geqslant b, \quad cx < f \tag{10}$$

et

$$uA = c, \quad u \geqslant 0, \quad ub \geqslant f \tag{11}$$

est résoluble quelles que soient la matrice-ligne c à n éléments et le nombre f.

DÉMONSTRATION. Supposons que le système (10) est incompatible. Alors, pour tout $f' < f$, est aussi incompatible le système

$$Ax \geqslant b, \quad -cx \geqslant -f'. \tag{12}$$

On peut appliquer le théorème 5 au système (12). Selon ce théorème, il existe une matrice-ligne $\bar{v} = \| v, w \|$ à $n+1$ éléments, telle que $\bar{v} \geqslant 0$, $vA - wc = 0$ et $vb - wf' > 0$.

Démontrons que $w > 0$. En effet, dans le cas contraire on aurait $vA = 0$ et $vb > 0$, d'où il résulterait que le système $Ax \geqslant b$ est incompatible.

Notons v' la matrice-ligne $w^{-1}v$ à n éléments. Il va de soi que

$$v' \geqslant 0, \quad v'A = c, \quad v'b > f'. \tag{13}$$

Ainsi, pour tout $f' < f$ le système (13) est compatible et il découle de la proposition 10 que le système (11) est compatible.

Il nous reste à démontrer que les systèmes (10) et (11) ne peuvent pas simultanément être compatibles. Il est aisé de le faire. S'ils sont tous deux compatibles et x et u sont des solutions correspondantes, on a

$$cx = uAx \geqslant ub \geqslant f,$$

ce qui contredit l'inégalité $cx < f$. Le théorème est complètement démontré.

Il a été déjà signalé qu'il est souvent nécessaire d'étudier des systèmes d'inéquations de la forme (1) où l'on exige que les variables soient positives. Donnons l'énoncé du théorème de Farkas pour ce cas-là.

Proposition 13. *De deux systèmes d'inéquations linéaires*

$$Ax \geqslant b, \quad x \geqslant 0, \quad cx < f \tag{14}$$

et

$$uA \leqslant c, \quad u \geqslant 0, \quad ub \geqslant f \tag{15}$$

un système et un seul est obligatoirement compatible.

Comme dans le cas de la proposition 12, considérons la matrice $\bar{A}$ et la matrice-colonne $\bar{b}$ permettant d'écrire le système $Ax \geqslant b$, $x \geqslant 0$ sous la forme $\bar{A}x \geqslant \bar{b}$. En appliquant le théorème 6, on constate que le système $\bar{A}x \geqslant \bar{b}$, $cx < f$ est compatible si et seulement si est incompatible le système

$$uA + v = c, \quad u \geqslant 0, \quad v \geqslant 0, \quad ub \geqslant f,$$

qu'il n'est pas difficile de réduire à la forme (15).

Appliquons le théorème de Farkas à un cas particulier quand le système (1) est un système d'équations linéaires. Ecrivons ce système comme réunion des systèmes $Ax \geqslant b$ et $-Ax \geqslant -b$ et démontrons la proposition qui suit.

Proposition 14. *L'inéquation $cx \geqslant f$ est une conséquence du système compatible d'équations linéaires $Ax = b$ si et seulement si le système*

$$\begin{aligned} Ax &= b \\ cx &\geqslant f \end{aligned} \tag{16}$$

est compatible et la matrice-ligne c est une combinaison linéaire des lignes de la matrice A.

Remarquons que la formulation géométrique de cette proposition est assez évidente : un plan $(n - r)$-dimensionnel est contenu dans le demi-espace donné si et seulement s'il y possède au moins un point et s'il est parallèle à l'hyperplan frontière de ce demi-espace.

DÉMONSTRATION. 1. Supposons que $\boldsymbol{cx} \geqslant f$ résulte du système $A\boldsymbol{x} = \boldsymbol{b}$. Il est clair que le système (16) est compatible. En outre, selon le théorème 6, il existe des matrices-lignes positives $\boldsymbol{u}$ et $\boldsymbol{v}$ pour lesquelles $\boldsymbol{c} = \boldsymbol{u}A - \boldsymbol{v}A$ et $\boldsymbol{ub} - \boldsymbol{vb} \leqslant f$. En particulier, cela signifie que $\boldsymbol{c} = (\boldsymbol{u} - \boldsymbol{v})A$, c'est-à-dire que $\boldsymbol{c}$ est une combinaison linéaires des lignes de A.

2. Supposons que $\boldsymbol{c} = \boldsymbol{p}A$ et que le système (16) est compatible. Posons

$$u_i = p_i, \quad v_i = 0 \quad \text{si} \quad p_i \geqslant 0,$$

et

$$u_i = 0, \quad v_i = -p_i \quad \text{si} \quad p_i < 0, \quad i = 1, \ldots, m.$$

Ainsi, $\boldsymbol{u} \geqslant \boldsymbol{0}$, $\boldsymbol{v} \geqslant \boldsymbol{0}$ et $\boldsymbol{p} = \boldsymbol{u} - \boldsymbol{v}$. Pour la matrice-ligne positive $\| \boldsymbol{u}, \boldsymbol{v}, 1 \|$ à $2m + 1$ éléments, il ressort de

$$\| \boldsymbol{u}, \boldsymbol{v}, 1 \| \cdot \begin{Vmatrix} -A \\ A \\ \boldsymbol{c} \end{Vmatrix} = 0$$

en vertu de la compatibilité du système (16) que

$$\| \boldsymbol{u}, \boldsymbol{v}, 1 \| \cdot \begin{Vmatrix} -\boldsymbol{b} \\ \boldsymbol{b} \\ f \end{Vmatrix} \leqslant 0,$$

autrement dit, $\boldsymbol{ub} - \boldsymbol{vb} \geqslant f$. Joint à $\boldsymbol{p}A = \boldsymbol{u}A - \boldsymbol{v}A = \boldsymbol{c}$, $\boldsymbol{u} \geqslant \boldsymbol{0}$, $\boldsymbol{v} > \boldsymbol{0}$, cela signifie que l'inéquation $\boldsymbol{cx} \geqslant f$ se déduit du système $A\boldsymbol{x} \geqslant \boldsymbol{b}$, $-A\boldsymbol{x} \geqslant -\boldsymbol{b}$, ce qu'il fallait démontrer.

6. Principe des solutions frontières. On démontrera ici le théorème qui porte le nom de principe des solutions frontières.

THÉORÈME 7. *Supposons que le système* (1) *est compatible et* Rg $A = r \geqslant 1$. *Il existe alors une sous-matrice A' composée de r lignes de A, telle que* Rg $A' = r$ *et que $A'\boldsymbol{x} = \boldsymbol{b}'$ entraîne $A\boldsymbol{x} \geqslant \boldsymbol{b}$, où $\boldsymbol{b}'$ est la matrice-colonne des éléments de* $\boldsymbol{b}$ *qui correspondent aux lignes de A'.*

Du point de vue géométrique, le théorème signifie que l'ensemble polyédrique convexe des solutions du système (1) admet au moins une face qui est un plan de dimension $n - r$.

La démonstration consiste à choisir successivement les lignes qui forment la sous-matrice A'.

1. Montrons que si $r \geqslant 1$, le système compatible de la forme (1) admet une solution pour laquelle l'une au moins de ses inéquations devient égalité. Soit $\boldsymbol{x}_0$ la solution du système pour laquelle aucune de ses inéquations ne devienne égalité. Vu que $r \geqslant 1$, une au moins des inéquations ne se vérifie pas identiquement et, par suite, il existe une matrice-colonne $\boldsymbol{x}_1$ ne véri-

fiant pas le système. Considérons le segment joignant x_0 et x_1 :

$$x_t = x_0 + t(x_1 - x_0), \qquad t \in [0, 1]. \tag{17}$$

Un calcul simple montre que l'inéquation $a_i x_t \geqslant b_i$, où a_i est la i-ième ligne de A, est équivalente à

$$t \leqslant \frac{1}{1 + \alpha_i^2},$$

où

$$\alpha_i^2 = - \frac{a_i x_1 - b_i}{a_i x_0 - b_i}$$

pour $a_i x_1 - b_i < 0$ et $\alpha_i = 0$ pour $a_i x_1 - b_i \geqslant 0$. Soit

$$\tau = \min_i \left\{ \frac{1}{1 + \alpha_i^2} \right\} = \frac{1}{1 + \alpha_{i_1}^2}.$$

Il est alors aisé de vérifier que x_τ est solution du système (1) et que $a_{i_1} x_\tau = b_{i_1}$.

Ainsi, le système

$$a_{i_1} x = b_{i_1}, \tag{18}$$

$$a_i x \geqslant b_i, \qquad i \in \{1, m\}, \qquad i \neq i_1 \tag{19}$$

est compatible. Si $r = 1$, chacune des inéquations (19) est en vertu de la proposition 14 une conséquence de l'équation (18), ce qui prouve le théorème. Dans le cas contraire, il existe une inéquation de numéro i_2 ne se vérifiant pas pour une solution de l'équation (18).

2. Supposons maintenant qu'on a choisi $k < r$ lignes de A de numéros $i_1, \ldots, i_k$, de manière que ces lignes soient linéairement indépendantes et le système

$$a_i x = b_i, \qquad i \in \{i_1, \ldots, i_k\} = I, \tag{20}$$

$$a_j x \geqslant b_j, \qquad j \notin I, \tag{21}$$

soit compatible. En vertu de la proposition 14, une au moins des inéquations (21) ne se déduit pas de (20) et par suite, il existe un x_1 tel que $a_i x_1 = b_i$, $i \in I$, et pour un j on a $a_j x_1 < b_j$.

Soit x_0 la solution du système (20), (21). Considérons le segment défini par la formule (17). Vu que x_0 et x_1 vérifient (20), il en est de même de x_t pour tout t. Par analogie à la première partie de la démonstration, on peut choisir τ sur le segment [0, 1] de manière que x_τ soit la solution du système (21) pour laquelle l'une de ses inéquations devient égalité. Soit i_{k+1} le numéro de cette inéquation.

Remarquons que pour les inéquations qui se déduisent du système d'équations (20) on a obligatoirement $a_l x_1 - b_l \geqslant 0$ et, par suite, $\alpha_l^2 = 0$. Or la valeur minimale de τ correspond au maximum de α_j qui est positif. C'est pourquoi la ligne $a_{i_{k+1}}$ n'est pas une combinaison linéaire des lignes a_i, $i \in I$.

L'adjonction de l'égalité $a_{i_{k+1}} x = b_{i_{k+1}}$ au système (20) et l'exclusion de l'inéquation correspondante de (21) permettent d'obtenir le système compatible de la forme (20), (21), contenant $k + 1$ égalités linéairement indépendantes.

Maintenant, si $k + 1 = r$, le théorème est démontré. Mais si $k + 1 < r$, la procédure de choix de lignes peut être continuée, ce qui achève la démonstration.

On a vu dans la proposition 6 que l'ensemble polyédrique convexe ne peut avoir de faces de dimension strictement inférieure à $n - r$. D'autre part, on va démontrer que l'ensemble des solutions du système d'inéquations linéaires (1) ne peut contenir de plan dont la dimension est strictement supérieure à $n - r$. La traduction algébrique de ce fait est la proposition suivante :

PROPOSITION 15. *Supposons que toutes les inéquations du système* (1) *sont vérifiées par les solutions du système compatible d'équations linéaires*

$$c_j x = f_j, \quad j = 1, \ldots, s, \tag{22}$$

les matrices-lignes $c_1, \ldots, c_s$ *étant linéairement indépendantes. Alors* $s \geqslant \operatorname{Rg} A$.

En effet, il découle de la proposition 14 que chaque ligne de la matrice A est une combinaison linéaire des matrices-lignes $c_1, \ldots, c_s$. Donc, $s \geqslant \operatorname{Rg} A = r$ et la dimension $n - s$ du plan défini par les équations (22) est inférieure à $n - r$.

Ainsi, le théorème 7 démontre l'existence d'une face plane de dimension $n - r$; cette dimension est minimale parmi les dimensions des faces et maximale parmi celles des plans contenus dans l'ensemble polyédrique.

§ 3. Eléments de programmation linéaire

1. Introduction. Les méthodes de recherche de la plus grande (ou de la plus petite) valeur d'une fonction linéaire sur un ensemble polyédrique convexe sont réunies sous la dénomination commune de programmation linéaire.

La programmation linéaire, comme branche autonome, est apparue à la fin des années 40 — début des années 50, c'est-à-dire à peu près à la même époque que les premiers ordinateurs, et s'est développée grâce essentiellement au perfectionnement des techniques numériques. Le besoin

d'appliquer l'appareil mathématique et les possibilités des ordinateurs à un large domaine de problèmes pratiques a engendré la création de modèles d'optimisation linéaires en économie, technique, médecine, domaine militaire, etc.

Le mot « modèle » signifie dans ce contexte une description approchée de la situation réelle et du problème associé, laquelle conserve toutefois une ressemblance qui doit être suffisante pour les objectifs pratiques. Dans les modèles mathématiques, la description se fait à l'aide de l'appareil et des symboles mathématiques. En particulier, les modèles linéaires utilisent des fonctions linéaires et des systèmes d'équations ou d'inéquations linéaires. Les modèles d'optimisation se distinguent par le fait qu'ils impliquent la résolution d'un problème d'extrémum lié.

Les principaux problèmes envisagés dans la programmation linéaire relèvent du domaine économique. Pour simplifier l'exposé on ne parlera que d'eux. Pour fixer les idées, étudions le modèle suivant appelé *interprétation économique standard* du problème général de programmation linéaire.

On considère une entreprise qui a n « activités » faisant chacune intervenir un certain nombre de m « biens » (par exemple, heures-ouvriers, machines-outils, énergie électrique, matériaux ...). On sait que pour chacune des n activités j il faut a_{ij} unités du bien i. On connaît aussi la quantité disponible de chaque bien i : b^i, $i = 1, \ldots, m$, et la valeur du profit tiré de la vente d'une unité de chacune des n activités : c_j, $j = 1, \ldots, n$. Il s'agit d'indiquer combien d'unités de chaque activité on doit produire pour obtenir le profit maximal.

Si l'entreprise produit x^j unités se rapportant à l'activité j, $j = !, \ldots, n$, elle dépensera

$$a_{i1}x^1 + \ldots + a_{in}x^n$$

unités du bien i. Cette dépense ne doit pas excéder la disponibilité b^i. Ainsi, les limitations imposées aux biens prennent la forme d'un système d'inéquations linéaires

$$a_{i1}x^1 + \ldots + a_{in}x^n \leqslant b^i, \quad i = 1, \ldots, m.$$

Le profit total engendré par toutes les activités de l'entreprise sera égal à

$$\varphi(x) = c_1x^1 + \ldots + c_nx^n.$$

On doit déterminer $x^1, \ldots, x^n$ de telle sorte qu'ils vérifient le système d'inéquations linéaires et maximisent la fonction φ.

Dès la création des premiers modèles d'optimisation linéaires on s'est aperçu que le problème de programmation linéaire correspondant pouvait être entièrement résolu. Ceci a entraîné une intense prolifération des

recherches en la matière et un aussi intense développement de la programmation linéaire qui était confrontée à de nouveaux problèmes.

La majorité des travaux consacrés à la programmation linéaire se fixaient pour objectif de résoudre des problèmes de dimension aussi grande que possible (c'est-à-dire à un plus grand nombre de variables et de contraintes) et aussi rapidement que possible. L'analyse de la précision de la solution obtenue passait au second plan. Cette situation était due non seulement au fait qu'on remettait à plus tard la résolution des problèmes laborieux et sans rendement immédiat, mais aussi au fait que la source principale d'erreurs dans les problèmes économiques résidait dans l'inadéquation du modèle, c'est-à-dire sa non-conformité à la réalité, ainsi que dans un mauvais recueil de l'information de départ. Même si, en s'accumulant, les erreurs d'arrondi pouvaient conduire à des solutions fantaisistes, elles n'inquiétaient pas les économistes qui les négligeaient devant toutes les autres erreurs.

Somme toute, on peut dire que la programmation linéaire est devenue en peu de temps une partie intégrante d'une nouvelle science, l'économie mathématique, et une branche importante des mathématiques appliquées, riche en idées, acquis et liens avec ses autres branches.

On présentera dans ce chapitre les principaux acquis de la programmation linéaire et quelques-unes de ses applications.

2. Position du problème. Dans ce paragraphe, on suppose qu'est donné un système d'inéquations linéaires de la forme

$$x \geqslant \mathbf{0}, \quad Ax \geqslant b, \tag{1}$$

ou sous une forme plus développée

$$\begin{gathered} a_{11}x^1 + \ldots + a_{1n}x^n \geqslant b_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_{m1}x^1 + \ldots + a_{mn}x^n \geqslant b_m, \\ x^1 \geqslant 0, \ldots, x^n \geqslant 0, \end{gathered}$$

et l'on recherche la solution $x^1, \ldots, x^n$ de ce système sur laquelle la fonction

$$\varphi(x) = c_1x^1 + \ldots + c_nx^n \tag{2}$$

prend une valeur minimale.

La fonction (2) est appelée *fonction économique* du problème, le système d'inéquations (1) *système de contraintes,* toute solution du système d'inéquations (1) est dite *point admissible* et le point où la fonction présente sa valeur minimale, *solution* du problème. Il faut remarquer qu'il existe plusieurs systèmes de termes pour décrire les problèmes de programmation linéaire.

Comme au § 2, pour une interprétation géométrique du système d'iné-

quations (1) on se servira de l'espace affine. Pour le but poursuivi, on peut se limiter à un repère choisi une fois pour toutes, ce qui en fait transforme l'espace affine en espace arithmétique. On identifiera tout point à sa colonne de coordonnées.

Le problème qu'on vient de formuler n'est pas un problème de forme générale mais il possède une propriété permettant de réduire tout problème de programmation linéaire à cette forme à l'aide de transformations pas trop compliquées.

Premièrement, il se peut qu'il faille trouver le maximum et non pas le minimum de la fonction. Or il est évident que si on sait trouver le minimum, on arrivera à obtenir le maximum, car le maximum de la fonction $\varphi(x)$ est atteint au point où est atteint le minimum de la fonction $-\varphi(x)$.

Deuxièmement, le système d'inéquations linéaires (1) contient les contraintes de positivité des variables. On a vu à la fin du § 1 que tout système d'inéquations linéaires peut être transformé en un système à variables positives si au lieu de chacune des variables initiales x^i on introduit deux variables positives y^i et z^i telles que $x^i = y^i - z^i$. On parlera plus en détail des transformations du problème de programmation linéaire au point 2 du § 4.

Les contraintes imposées aux variables entraînent que le rang du système d'inéquations (1) est obligatoirement égal à n et, par suite, l'ensemble polyédrique convexe des points admissibles possède des sommets s'il n'est évidemment pas vide.

Le problème de programmation linéaire n'est pas obligé d'avoir une solution. Il peut arriver que le système de contraintes ne soit pas compatible. Même un système de contraintes compatible peut ne pas avoir de solution si la fonction économique n'est pas minorée. Ce dernier cas n'est possible que si l'ensemble polyédrique n'est pas borné.

Dans ce paragraphe, on ne s'attardera pas sur les procédés de résolutions du problème et l'on étudiera ses propriétés fondamentales.

3. Existence de solution. Avant tout, on établira les conditions sous lesquelles le problème

$$x \geqslant 0, \quad Ax \geqslant b, \quad cx \to \min \tag{3}$$

est résoluble et on décrira l'ensemble des solutions s'il n'est pas vide.

THÉORÈME 1. *Si le système des contraintes* (1) *est compatible et la fonction économique est minorée, le problème* (3) *est résoluble. La fonction* φ *présente son minimum en l'un des sommets de l'ensemble polyédrique* $\mathscr{A}$ *défini par le système* (1) (*et peut-être en d'autres points*).

Pour démontrer ce théorème, représentons un point arbitraire de l'ensemble, en accord avec le théorème 1 du § 2, sous la forme

$$x = u + v.$$

Ici

$$u = \sum_{i=1}^{N} \alpha_i x_i$$

est une combinaison convexe des sommets $x_1, \ldots, x_N$ de l'ensemble $\mathscr{A}$, *et*

$$v = \sum_{j=1}^{M} \beta_j l_j$$

un vecteur du cône dont la carcasse est $l_1, \ldots, l_M$.

Notons avant tout que $cl_j \geqslant 0$ pour tous les $j = 1, \ldots, M$. En effet, s'il existait un vecteur l_k pour lequel $cl_k < 0$, on pourrait en augmentant β_k, les autres coefficients étant fixés, diminuer indéfiniment la valeur de la fonction

$$\varphi(x) = cx = \beta_k cl_k + \sum_{j \neq k} \beta_j cl_j + \sum_i \alpha_i cx_i,$$

ce qui contredit l'hypothèse que φ est minorée.

Associons au point x le point x' déduit de x par substitution de zéro à tous les coefficients β_j dans sa décomposition. Alors

$$cx \geqslant cx',$$

car cx' est déduit de cx par élimination de termes positifs. Notons φ_0 le plus petit des nombres $cx_1, \ldots, cx_N$. On a alors l'estimation suivante :

$$cx' = \sum_{i=l}^{N} \alpha_i cx_i \geqslant \varphi_0 \sum_{i=1}^{N} \alpha_i = \varphi_0.$$

Ainsi, tout point admissible x vérifie l'inégalité

$$cx \geqslant \varphi_0,$$

où φ_0 est une valeur de la fonction $\varphi(x) = cx$ en l'un des sommets, ce qui achève la démonstration.

On dira que le point x est un *point intérieur* à l'ensemble polyédrique convexe $\mathscr{A}$ défini par le système d'inéquations (1) si x vérifie toutes les contraintes-inégalités du système en tant qu'inégalités strictes.

PROPOSITION 1. *Si la fonction linéaire $\varphi(x)$ n'est pas constante sur l'ensemble polyédrique $\mathscr{A}$, elle ne peut atteindre la valeur minimale en aucun point intérieur à cet ensemble.*

Soit ρ le rang du système formé à partir des contraintes-égalités du système (1). Selon la proposition 5 du § 2, l'ensemble polyédrique est situé dans le plan $(n - \rho)$*-dimensionnel* $\mathscr{S}_{n-\rho}$ défini par les contraintes-égalités.

Supposons que les inéquations sont numérotées de façon que $a^{m-\rho+1}$, ..., a^m sont les matrices-lignes des coefficients de ces contraintes. On peut résoudre le système d'équations

$$a^j x = b^j, \quad j = m - \rho + 1, \ldots, m,$$

par rapport à ρ variables. Posons que les numéros de ces variables sont $n - \rho + 1, \ldots, n$. Portons ces variables dans les inéquations de numéros $\leqslant m - \rho$ et dans l'expression de la fonction économique. Dans ce cas, les contraintes-égalités restantes (linéairement dépendantes) sont vérifiées identiquement, tandis que les contraintes-inégalités prennent la forme

$$\bar{a}^i \bar{x} \geqslant \bar{b}^i,$$

où $x = {}^t\|x^1, \ldots, x^{n-\rho}\|$.

La restriction de la fonction économique au plan $\mathscr{S}_{n-\rho}$ est une fonction linéaire

$$\bar{\varphi}(\bar{x}) = \bar{c}\bar{x} + f$$

à terme constant f. La fonction $\bar{\varphi}$ est constante sur $\mathscr{A}$ si et seulement si la matrice-ligne de coefficients $\bar{c} = \|\bar{c}_1, \ldots, \bar{c}_{n-\rho}\|$ est nulle. Aussi, considère-t-on que $\bar{c} \neq \mathbf{0}$.

Soit x_0 un point intérieur à l'ensemble polyédrique $\mathscr{A}$. Compte tenu des contraintes-égalités, le point x_0 est défini par la matrice-colonne $\bar{x}_0$ composée des $n - \rho$ *premiers éléments de la matrice-colonne* x_0. Considérons la demi-droite

$$\bar{x} = \bar{x}_0 - {}^t\bar{c}t,$$

où t est un paramètre à valeurs positives. Pour tout $i = 1, \ldots, m - \rho$, il vient $\bar{a}^i\bar{x}_0 > \bar{b}^i$ et

$$\bar{a}^i\bar{x} = \bar{a}^i\bar{x}_0 - \bar{a}^i({}^t\bar{c})t.$$

Si $\bar{a}^i({}^t\bar{c}) \leqslant 0$, tous les points de la demi-droite vérifient l'inéquation $\bar{a}^i\bar{x} > \bar{b}^i$. Mais si $\bar{a}^i({}^t\bar{c}) > 0$, l'inéquation $\bar{a}^i\bar{x} \geqslant 0$ est vérifiée pour les points de la demi-droite auxquels correspondent les valeurs du paramètre $t \leqslant t_i$, où

$$t_i = \frac{\bar{a}^i\bar{x}_0}{\bar{a}^i({}^t\bar{c})}.$$

Il va de soi que tous les $t_i > 0$. Supposons qu'au point $\bar{x}$ correspond une valeur strictement positive du paramètre, inférieure à tous les t_i. *Ce point appartient à* $\mathscr{A}$, c'est-à-dire est un point admissible.

En outre, pour $t > 0$

$$\bar{c}(\bar{x}_0 - {}^t\bar{c}t) = \bar{c}\,\bar{x}_0 - \bar{c}{}^t\bar{c}t < \bar{c}\,\bar{x}_0,$$

et, par suite, le point $\bar{x}_0$ n'est pas un point de minimum de $\bar{c}\,\bar{x} + f$, et donc celui de la fonction φ.

Il ressort de la proposition qu'on vient de démontrer que la valeur minimale ne peut être atteinte qu'en un point d'une face $\mathscr{A}'$ de l'ensemble $\mathscr{A}$. Or chaque face est aussi un ensemble polyédrique convexe auquel s'applique la même proposition : si la fonction n'est pas constante sur la face $\mathscr{A}'$, la valeur minimale ne peut être atteinte qu'en un point de sa face $\mathscr{A}''$. En appliquant successivement ces raisonnements aux faces de dimensions de plus en plus petites, on aboutit au théorème suivant.

THÉORÈME 2. *Les solutions du problème de programmation linéaire* (3) *remplissent une face de l'ensemble polyédrique défini par le système de contraintes. La solution est unique si et seulement si cette face est un sommet.*

En vertu des inégalités $x^i \geqslant 0$, le rang de la matrice du système (1) est n et l'ensemble polyédrique défini par le système (1) admet des sommets. Il en est donc de même de toute face de cet ensemble, de sorte que si le problème est résoluble, une de ses solutions est toujours un sommet, comme on l'a affirmé au théorème 1.

4. Problème dual. On peut rattacher au problème de programmation linéaire (3) un autre problème, appelé *problème dual*. Il consiste à trouver le maximum de la fonction linéaire

$$\psi(u) = ub,$$

définie sur l'espace des matrices-lignes à m éléments, à condition que l'argument u appartienne à un ensemble polyédrique convexe défini par les contraintes

$$u \geqslant 0, \quad uA \leqslant c. \tag{4}$$

Ainsi, le problème dual peut être écrit sous la forme suivante :

$$u \geqslant 0, \quad uA \leqslant c, \quad ub \to \max. \tag{5}$$

La relation entre les problèmes (3) et (5) est une équivalence, c'est-à-dire que le problème dual de (5) est équivalent au problème (3). En effet, pour construire le dual du problème (5), ce dernier doit être réduit à la forme (3). Pour le faire, multiplions la matrice A, la matrice-colonne b et la matrice-ligne c par -1 et écrivons u sous la forme de matrice-colonne ${}^t u$. Il vient

$$ {}^t u \geqslant 0, \quad -{}^t A\, {}^t u \geqslant -{}^t c, \quad -{}^t b\, {}^t u \to \min \tag{5'}$$

(la recherche du minimum de $-\psi(u)$ est équivalente à la recherche du maximum de $\psi(u)$). Le dual de ce problème prend la forme

$$y \geqslant 0, \quad y(-{}^t A) \leqslant -{}^t b, \quad y(-{}^t c) \to \max. \tag{3'}$$

En remplaçant y par ${}^t x$ et en changeant les signes de $-{}^t A$, $-{}^t b$ et $-{}^t c$, on réduit (3') à la forme (3).

PROPOSITION 2. *Si x et u sont des points admissibles des problèmes* (3) *et* (5), *il vient*

$$ub \leqslant cx. \tag{6}$$

De plus, si pour des points admissibles x_0 et u^0 est vérifié $cx_0 = u^0 b$, alors x_0 et u^0 sont les solutions des problèmes (3) *et* (5) *respectivement.*

DÉMONSTRATION. Etant donné que x et u sont positifs, on a

$$ub \leqslant uAx \leqslant cx.$$

En outre, pour tout point admissible du problème (3), l'inégalité démontrée entraîne que

$$cx \geqslant u^0 b = cx_0.$$

Aussi, cx_0 est-elle la plus petite des valeurs de cx sur les points admissibles. D'une façon analogue, pour tout point admissible du problème (5),

$$ub \leqslant cx_0 = u^0 b.$$

Cela signifie que $u^0 b$ est la plus grande des valeurs de la fonction ub sur les points admissibles.

Le théorème suivant porte le nom de *théorème de dualité*.

THÉORÈME 3. *Si le problème primal* (3) *possède une solution x_0, le problème dual* (5) *possède alors une solution u^0, de sorte que $cx_0 = u^0 b$.*

Si la fonction économique dans le problème primal n'est pas minorée, le système des contraintes du problème dual est incompatible.

La démonstration se fonde sur la proposition 13 du § 2, selon laquelle un et un seul des deux systèmes d'inéquations

$$x \geqslant 0, \quad Ax \geqslant b, \quad cx < f \tag{7}$$

et

$$u \geqslant 0, \quad uA \leqslant c, \quad ub \geqslant f \tag{8}$$

est compatible.

Le fait que la fonction cx n'est pas minorée sur l'ensemble polyédrique (1) signifie que le système (7) est compatible pour tout f, et, par suite, le système (8) est incompatible, quel que soit f. Admettons que le problème (5) possède un point admissible u et posons $f = ub$. Dans ce cas, u est solution du système (8) pour ce f, ce qui contredit l'hypothèse. Ainsi, le problème (5) n'a pas de point admissible, et la seconde assertion du théorème est démontrée.

Démontrons la première assertion. Si le problème (3) a une solution x_0, c'est-à-dire si cx_0 est la valeur minimale de la fonction cx sur l'ensemble polyédrique (1), le système (7) est incompatible pour $f < cx_0$, et, par suite, le système (8) est compatible pour ces valeurs de f. Inversement, pour $f \geqslant cx_0$ le système (7) est compatible et par suite, (8) est incompatible. On en déduit que les systèmes des contraintes des deux problèmes sont compatibles. Il est de plus essentiel que pour chacun des problèmes (7) et (8) il existe une valeur f pour laquelle ce système est compatible.

Partageons l'ensemble des nombres réels en deux classes $\mathscr{U}$ et $\mathscr{V}$. Rapportons à la classe $\mathscr{U}$ les nombres f pour lesquels le système (7) est compatible, et à la classe $\mathscr{V}$ ceux pour lesquels est compatible le système (8). On a vu que les deux classes ne sont pas vides et chaque nombre appartient à une classe et une seule. Démontrons que tout nombre de la classe $\mathscr{U}$ est plus grand que tout nombre de la classe $\mathscr{V}$. En effet, soit $f' \leqslant f''$ et pour $f = f'$ le système (7) est compatible, tandis que pour $f = f''$ est compatible le système (8). Or $cx < f'$ entraîne $cx < f''$ et, par suite, le système (7) est aussi compatible pour $f = f''$, ce qui contredit l'hypothèse.

Selon le principe de Dedekind, la partition réalisée définit un seul nombre f_0 qui est inférieur à tout nombre de la classe $\mathscr{U}$ et est supérieur à tout nombre de la classe $\mathscr{V}$. Quant au nombre f_0, il peut, en général, appartenir à l'une quelconque des deux classes. Etudions les deux possibilités.

Soit $f_0 \in \mathscr{V}$. Dans ce cas, il existe un point u^1 tel que

$$u^1 \geqslant 0, \quad u^1 A \leqslant c, \quad u^1 b \geqslant f_0.$$

Si $u^1 b > f_0$, alors $u^1 b \in \mathscr{U}$ et il existe un point x tel que

$$x \geqslant 0, \quad Ax \geqslant b, \quad cx < u^1 b.$$

Cela est contraire à l'inégalité (6). Donc, $u^1 b = f_0$.

D'autre part, le système (7) est incompatible pour f_0 et par suite, la solution x_1 du problème (3) vérifie $cx_1 \geqslant f_0$. Démontrons qu'on est ici obligatoirement en présence d'une égalité.

En effet, soit $cx_1 > f_0$. Alors $cx_1 \in \mathscr{U}$ et le système

$$x \geqslant 0, \quad Ax \geqslant b, \quad cx < cx_1$$

est compatible. Cela signifie que la valeur de la fonction cx_1 n'est pas minimale. Ainsi, on a montré que

$$u^1 b = f_0 = cx_1.$$

Considérons maintenant le cas de $f_0 \in \mathscr{U}$. Il existe alors un point x_1 pour lequel

$$x_1 \geqslant 0, \quad Ax_1 \geqslant b, \quad cx_1 < f_0.$$

La dernière inégalité montre que $cx_1 \in \mathcal{V}$ et, par suite, il existe un point u^1 vérifiant le système d'inéquations :

$$u^1 \geqslant 0, \quad u^1 A \leqslant c, \quad u^1 b \geqslant cx_1.$$

Selon (6) la dernière inégalité ne peut avoir lieu que si est vérifiée l'égalité

$$u^1 b = cx_1.$$

Ainsi, quelle que soit la classe $\mathcal{U}$ ou $\mathcal{V}$ qui contient le nombre f_0, il existe des points x_1 et u^1 tels que $cx_1 = u^1 b$. Il découle de la proposition 2 qu'ils sont des solutions des problèmes (3) et (5) respectivement. Ensuite, il est évident que quelles que soient les solutions x_0 et u^0, on a les égalités $cx_1 = cx_0$ et $u^1 b = u^0 b$. Il en découle que pour toutes solutions se vérifie l'égalité $cx_0 = u^0 b$, ce qui achève la démonstration du théorème.

En vertu de la réciprocité du lien entre les systèmes duals, il ressort du théorème 3 que si la fonction ub n'est pas majorée sur l'ensemble polyédrique (4), le système des contraintes (1) du problème (3) est incompatible.

Il faut remarquer que la réciproque de cette assertion (et de la deuxième assertion du théorème 1) n'est pas vraie. L'incompatibilité du système des contraintes d'un des problèmes n'entraîne pas que la fonction économique du problème dual n'est pas bornée : les deux systèmes de contraintes peuvent être incompatibles comme le montre l'exemple suivant.

Le problème dual de

$$x^1 \geqslant 0, \quad x^2 \geqslant 0, \quad \begin{matrix} x^1 - 2x^2 \geqslant 1, \\ -x^1 + 2x^2 \geqslant 0, \end{matrix} \quad -x^2 \to \min$$

est

$$u_1 \geqslant 0, \quad u_2 \geqslant 0, \quad \begin{matrix} u_1 - u_2 \leqslant 0, \\ -2u_1 + 2u_2 \leqslant -1, \end{matrix} \quad u_1 \to \max.$$

On voit facilement que dans chaque problème les contraintes sont incompatibles.

Considérons un corollaire important du théorème 3. Si x_0 et u^0 sont les solutions des problèmes duals (3) et (5), il découle du théorème 3 que

$$u^0 b = u^0 A x_0. \tag{9}$$

En effet, on a les relations $u^0 b \leqslant u^0 A x_0 \leqslant cx_0$ et $u^0 b = cx_0$. L'égalité (9) peut être mise sous la forme $u^0 y = 0$, où $y = Ax_0 - b$.

Les éléments de la matrice-ligne u^0 et de la matrice-colonne y sont positifs. Cela signifie que dans l'expression

$$u^0 y = u_1^0 y^1 + u_2^0 y^2 + \ldots + u_m^0 y^m$$

tous les termes non nuls sont strictement positifs. Donc, $u^0 y = 0$ entraîne que tous les termes sont égaux à zéro.

Soient a_k^i les éléments de la matrice A, et b^i et x_0^k les éléments des matrices-colonnes b et x_0. Alors,

$$y^i = \sum_{k=1}^{n} a_k^i x_0^k - b^i, \quad i = 1, \ldots, m,$$

et l'on aboutit à la proposition suivante.

PROPOSITION 3. *Si x_0 et u^0 sont solutions des problèmes duals* (3) *et* (5) *et si pour un numéro $i \in [1, m]$ est vérifié*

$$\sum_{k=1}^{n} a_k^i x_0^k > b^i,$$

la i-ième coordonnée u_i^0 du point u^0 est égale à zéro. Inversement, si $u_j^0 > 0$, on a

$$\sum_{k=1}^{n} a_k^j x_0^k = b^j.$$

Remarquons que l'annulation simultanée de u_j^0 et de y^i est possible.

De façon analogue on peut démontrer la

PROPOSITION 4. *Pour les solutions x_0 et u^0 du couple de problèmes duals, il découle de*

$$\sum_{i=1}^{m} u_i^0 a_j^i < c_i$$

que $x^i = 0$, et de $x^i > 0$, que

$$\sum_{i=1}^{m} u_i^0 a_j^i = c_j.$$

Supposons que les éléments de la matrice-colonne b dans (1) sont des variables indépendantes. Les études de ce genre sont essentielles pour la recherche de l'influence de variations des données du problème sur la solution. Les variations peuvent être dues aussi bien à la modification du problème qu'aux erreurs de mesure des données initiales.

La matrice-colonne b n'intervient pas dans le système des contraintes du problème dual et les sommets $u^0, \ldots, u^p$ de l'ensemble polyédrique des points admissibles de ce problème ne dépendent pas de b. Si l'on varie un peu b, la fonction économique modifiée du problème dual atteindra sa valeur maximale au même sommet u^0 que l'ancienne fonction. Montrons-le en estimant la variation possible de b.

Il nous faut que pour tous $i = 1, \ldots, p$ soit vérifiée l'inégalité

$$u^0(b + \Delta b) \geqslant u^i(b + \Delta b),$$

ou bien

$$(u^0 - u^i)b \geqslant (u^i - u^0)\Delta b.$$

Le premier membre est ici positif vu que u^0 est la solution du problème primal. En vertu de l'inégalité de Cauchi on a $(u^i - u^0)\Delta b \leqslant \| u^i - u^0 \| \times$ $\times \| \Delta b \|$. Aussi l'inégalité exigée se vérifiera-t-elle pour tous les i si

$$\| \Delta b \| \leqslant \min_i \frac{(u^0 - u^i)b}{\| u^0 - u^i \|}. \tag{10}$$

Il s'ensuit la

Proposition 5. *Si la variation de la matrice-colonne b ne dépasse pas le second membre de la formule* (10), *la solution u^0 du problème dual* (5) *demeure inchangée.*

En vertu de la dualité, il en découle l'invariance de la solution du problème (3) pour de petits changements des coefficients c de la fonction économique φ.

Servons-nous de la proposition 5 pour obtenir la propriété suivante de la solution du problème dual.

La valeur minimale cx_0 de la fonction cx du problème (3) dépend naturellement de b car l'ensemble de points sur lequel est recherché le minimum varie avec b. De plus, on connaît la forme de cette dépendance :

$$cx_0 = u^0 b.$$

On ne peut pas dire que cette dépendance est linéaire car pour une importante variation de b la matrice-ligne u^0 ne reste plus inchangée. Mais dans un voisinage de la matrice-colonne b (sauf un nombre fini de ces matrices-colonnes) la fonction est linéaire. Indiquons que l'analogue unidimensionnel de cette fonction est la fonction dont la courbe représentative est une ligne brisée.

Ceci pris en compte, on peut affirmer que

$$\frac{\partial (cx_0)}{\partial b^i} = u_i^0 \tag{11}$$

en des points où ces dérivées partielles existent.

Supposons qu'une composante u_i^0 de la solution u^0 du problème dual est nulle. L'égalité $u_i^0 = 0$ demeure vraie après la substitution à la matrice-colonne b de la matrice-colonne voisine b'. Donc, la dérivée partielle de cx_0 par rapport à b^i vaut zéro dans un voisinage de la valeur initiale b^i. Ainsi,

l'égalité $u_i^0 = 0$ signifie que cx_0 ne varie pas pour de petites modifications de b^i.

D'autre part, selon la proposition 3, u_i^0 s'annule si la solution x_0 du problème primal (3) vérifie la i-ième contrainte en tant qu'inégalité stricte. On a ainsi obtenu la

PROPOSITION 6. *La valeur minimale de la fonction φ ne varie pas pour de petites modifications du i-ième élément de la matrice-colonne des seconds membres si et seulement si est nulle la i-ième composante de la solution du problème dual. Il suffit pour cela que la i-ième contrainte se vérifie pour x_0 comme inégalité stricte.*

5. Fonction de Lagrange. Une autre interprétation de la solution du problème dual est liée à la fonction

$$L(x, u) = cx + u(b - Ax) \tag{12}$$

composée pour le problème (3). u est ici la matrice-ligne à m éléments, si bien que $L(x, u)$ est une fonction de $m + n$ variables. On l'appelle *fonction de Lagrange*. En regroupant les termes, on peut la représenter sous la forme

$$L(x, u) = ub + (c - uA)x, \tag{13}$$

d'où l'on voit que pour le couple de problèmes duals la fonction de Lagrange est la même.

PROPOSITION 7. *Soient x_0 et u^0 les solutions des problèmes duals* (3) *et* (5). *Alors le point (x_0, u^0) de $\mathscr{R}^{m+n}$ est le point-selle de la fonction de Lagrange, c'est-à-dire que tous $x \geqslant 0$ et $u \geqslant 0$ vérifient la double inégalité*

$$L(x_0, u) \leqslant L(x_0, u^0) \leqslant L(x, u^0). \tag{14}$$

DÉMONSTRATION. Pour tout u positif, $Ax_0 \geqslant b$ entraîne $u(b - Ax_0) \leqslant$ $\leqslant 0$. Or, en vertu de l'égalité (9), $u^0(b - Ax_0) = 0$. Donc,

$$cx_0 + u(b - Ax_0) \leqslant cx_0 + u^0(b - Ax_0),$$

et la première inégalité (14) est démontrée. La seconde inégalité se démontre de façon analogue si on utilise l'expression (13) et la proposition 4. De la démonstration donnée il ressort que

$$L(x_0, u^0) = cx_0 = u^0 b.$$

La fonction de Lagrange est étudiée en analyse mathématique (voir Koudriavtsev [21], t. II, p. 96) où elle est construite pour obtenir l'extrémum lié de la fonction $f(x)$, avec contraintes de la forme $\varphi_i(x) = 0$, $i = 1, \ldots, m$, et prend la forme

$$L(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i \varphi_i(x).$$

Les variables $\lambda_1, \ldots, \lambda_m$ sont dénommées *facteurs de Lagrange*.

Dans le cas considéré, les contraintes sont des inéquations et les variables sont obligées d'être positives. Le théorème étendant la méthode de Lagrange aux problèmes avec contraintes plus générales est appelé *théorème de Kuhn-Tucker*. On peut le trouver par exemple dans le livre de Karmanov [18].

La proposition 7 qu'on vient de démontrer est le théorème de Kuhn-Tucker pour le cas d'une fonction linéaire avec contraintes de la forme (1). On voit que les variables du problème dual jouent le rôle de facteurs de Lagrange par rapport au problème primal.

§ 4. Méthode du simplexe

1. Introduction. Dans ce paragraphe on fera connaissance avec la méthode du simplexe. C'est la méthode principale parmi toutes les méthodes finies de résolution des problèmes de programmation linéaire, c'est-à-dire les méthodes qui fournissent en principe un résultat exact en un nombre fini d'opérations. C'est précisément l'efficacité de la méthode du simplexe qui assura à la programmation linéaire une grande popularité.

L'idée principale de la méthode du simplexe est assez évidente. Selon le théorème 1 du § 3, le minimum de la fonction économique φ est atteint en l'un des sommets de l'ensemble polyédrique défini par le système des contraintes. D'un sommet à l'autre on peut passer suivant l'arête. Ceci étant, les valeurs de la fonction économique φ diminuent si la projection de son gradient sur l'arête (autrement dit, la dérivée de φ dans la direction de l'arête) est strictement négative. Supposons qu'on a trouvé un sommet d'où il part une arête telle que la projection sur elle du gradient soit strictement négative. Il y a deux possibilités à considérer :

a) L'arête est un segment. Dans ce cas, la seconde extrémité du segment est un sommet où φ prend une valeur plus petite qu'au sommet de départ.

b) L'arête est une demi-droite. Il existe alors sur l'arête des points où φ prend des valeurs aussi petites que l'on veut, de sorte que le problème est irrésoluble car la fonction n'est pas minorée.

Admettons que le problème a une solution. Vu qu'après le passage de chaque arête la valeur de φ diminue, on ne peut revenir au sommet déjà dépassé. Le nombre de sommets étant fini, on doit aboutir en un nombre fini d'étapes à un sommet où est atteint le minimum. Ce sommet se caractérise par le fait que les projections du gradient sur toutes les arêtes issues de ce sommet sont positives.

L'avantage important de la méthode du simplexe est que tous les calculs nécessaires à la réalisation du schéma décrit se réalisent facilement au cours de transformation d'une matrice suivant l'algorithme proche de la méthode de Gauss.

Des autres méthodes de programmation linéaires il convient de mentionner les méthodes itératives qui, ces derniers temps, se perfectionnent intensément et sont de plus en plus utilisées. Faute de place on ne s'arrêtera pas sur ces méthodes.

2. Forme canonique du problème. La méthode du simplexe est appliquée aux problèmes de forme dite *canonique*. Plus précisément, le système des contraintes doit être de la forme

$$\begin{array}{l} a_{11}x^1 + \dots + a_{1n}x^n = b^1, \\ \dots\dots\dots\dots\dots\dots \\ a_{m1}x^1 + \dots + a_{mn}x^n = b^m, \end{array} \qquad x^1 \geqslant 0, \dots, x^n \geqslant 0. \tag{1}$$

On suppose que dans ce cas les seconds membres sont positifs,

$$b^1 \geqslant 0, \dots, b^m \geqslant 0,$$

et les lignes de coefficients, linéairement indépendantes. De la dernière hypothèse il résulte en particulier que $m \leqslant n$.

Le système (1) contient m contraintes-égalités indépendantes et n inégalités (également indépendantes) de la forme $x^j \geqslant 0$. Ainsi donc, si le système est compatible, il définit un ensemble polyédrique convexe de dimension $\leqslant n - m$, qui possède des sommets.

D'une façon générale, le problème posé peut ne pas présenter de forme canonique, et la méthode du simplexe ne peut alors être appliquée directement. Toutefois, en augmentant le nombre de variables, on peut réduire tout problème à un problème équivalent de forme canonique. L'équivalence est ici comprise au sens que d'après la solution du problème canonique on peut indiquer la solution du problème initial.

PROPOSITION 1. *Tout problème de programmation linéaire peut être réduit au problème de forme canonique.*

DÉMONSTRATION. Posons que le système des contraintes du problème initial est de la forme

$$\begin{array}{l} a_{11}x^1 + \dots + a_{1n}x^n \geqslant b^1, \\ \dots\dots\dots\dots\dots\dots \\ a_{m1}x^1 + \dots + a_{mn}x^n \geqslant b^m. \end{array} \tag{2}$$

Transformons d'abord les inéquations en équations :

$$a_{i1}x^1 + \dots + a_{in}x^n - y^i = b^i, \quad i = 1, \dots, m,$$

où y^i sont des variables positives. (Si le système (2) contient l'inégalité de la forme $x^j \geqslant 0$, il n'est évidemment pas nécessaire de la réduire à $x^j - y^j = 0$, $y^i \geqslant 0$.) Multiplions ensuite par -1 chacune des égalités dont le second membre est négatif.

Dans le système obtenu, introduisons, au lieu de chacune des variables x^j non assujetties à la condition $x^j \geqslant 0$, deux nouvelles variables positives liées à celle-ci par l'égalité

$$x^j = z^j - z^{n+j}.$$

Maintenant, si l'on élimine les équations linéairement dépendantes, le système obtenu ne diffère du système (1) que par les notations des coefficients et des variables.

Pour simplifier l'écriture, posons que dans le système initial il n'y avait pas d'inéquations de la forme $x^j \geqslant 0$ et qu'on a joint m variables supplémentaires y^i. Soit ${}^t\|x^1, \ldots, x^n\|$ un point admissible du système initial. Posons $z^j = x^j$, $z^{n+j} = 0$ si $x^j \geqslant 0$, et $z^j = 0$, $z^{n+j} = -x^j$ si $x^j < 0$. Désignons les différences

$$(a_{i1}x^1 + \ldots + a_{in}x^n) - b^i$$

par y^i. Alors ${}^t\|z^1, \ldots, z^{2n}, y^1, \ldots, y^m\|$ est un point admissible du problème canonique. Inversement, d'après le point admissible du problème canonique il est aisé de construire le point admissible du problème initial. Aussi les systèmes de contraintes des deux problèmes sont-ils compatibles ou incompatibles simultanément.

Pour la fonction économique $\varphi(x) = c_1x^1 + \ldots + c_nx^n$ du problème initial définissons la fonction correspondante du problème canonique :

$$\bar{\varphi}(z) = c_1z^1 + \ldots + c_nz^n - c_1z^{n+1} - \ldots - c_{2n}z^{2n} + 0y^1 + \ldots + 0y^m.$$

Si au point $z = {}^t\|z^1, \ldots, z^{2n}, y^1, \ldots, y^m\|$ est associé le point

$$x = \mathbf{P}(z) = {}^t\|z^1 - z^{n+1}, \ldots, z^n - z^{2n}\|,$$

il est évident que $\bar{\varphi}(z) = \varphi(x)$.

Les ensembles de valeurs des deux fonctions coïncident et $\varphi(x)$ est minorée si et seulement s'il en est de même de $\bar{\varphi}(z)$. Si $\bar{\varphi}(z)$ présente son minimum au point z_0, la fonction $\varphi(x)$ prend au point $x_0 = \mathbf{P}(z_0)$ la même valeur qui est aussi minimale. La proposition est démontrée.

Il convient de noter qu'en recherchant une démonstration simple et générale on ne lésinait pas sur les variables supplémentaires. Il existe d'autres procédés de réduction du problème à la forme canonique sans que le nombre de variables augmente si fortement ; il y en a même d'autres où ce nombre diminue. En voici un exemple.

Posons que le système (2) a une matrice de coefficients de rang n et que les n premières inéquations possèdent des seconds membres linéairement indépendants. Introduisons les nouvelles variables

$$y^j = a_{j1}x^1 + \ldots + a_{jn}x^n - b^j, \quad j = 1, \ldots, n.$$

Dans ce cas, les n premières inéquations prennent la forme $y^1 \geqslant 0$, ...

..., $y^n \geqslant 0$. Dans les autres inéquations, remplaçons $x^1, \ldots, x^n$ par leurs expressions en nouvelles variables, puis transformons-les en équations linéairement indépendantes avec seconds membres positifs comme on l'a fait dans la démonstration de la proposition 1.

Il existe d'autres procédés de transformation des contraintes, mais la méthode utilisée dans la démonstration a pris une grande extension vu sa simplicité.

On supposera dans la suite de ce paragraphe qu'on a à résoudre le problème canonique de programmation linéaire .

$$Ax = b, \quad x \geqslant 0, \quad cx \to \min \tag{3}$$

sous les hypothèses

$$\operatorname{Rg} A = m, \quad b \geqslant 0.$$

3. Dual du problème canonique. Pour définir le dual du problème (3), on mettra (3) sous la forme

$$x \geqslant 0, \quad Px \geqslant q, \quad cx \to \min,$$

où la matrice P et la matrice-colonne q ont la structure en blocs suivante :

$$P = \left\| \begin{matrix} A \\ -A \end{matrix} \right\|, \quad q = \left\| \begin{matrix} b \\ -b \end{matrix} \right\|.$$

Le problème dual s'écrit alors selon (5) du § 2 sous la forme

$$yP \leqslant c, \quad y \geqslant 0, \quad yb \to \max,$$

où y est une matrice-ligne à $2n$ éléments qu'on écrira comme réunion de deux matrices-lignes y^1 et y^2 à n éléments chacune : $y = \| y^1, y^2 \|$. Ceci permet d'écrire le problème dual sous une forme plus développée :

$$(y^1 - y^2)A \leqslant c, \quad y \geqslant 0, \quad (y^1 - y^2)b \to \max.$$

Outre la condition de positivité, les matrices-lignes y^1 et y^2 n'y figurent que sous la forme de la différence $y^1 - y^2$ qu'on notera u. La matrice-ligne u ne doit pas nécessairement être positive. Si $\| y^1, y^2 \|$ vérifie le système des contraintes, u vérifie le système $uA \leqslant c$.

Inversement, si pour une matrice-ligne u on a $uA \leqslant c$, il est aisé de construire une matrice-ligne y à $2n$ éléments, telle que $yP \leqslant c$ et $y \geqslant 0$. Par conséquent, le dual du problème canonique (3) se réduit au problème

$$uA \leqslant c, \quad ub \to \max, \tag{4}$$

qu'on appellera *problème dual* de (3).

Le problème (4) n'est pas canonique. On lui donnera une forme plus proche de la forme canonique si l'on introduit n variables positives $z^1, \ldots$ ..., z^n écrites en matrice-colonne z et l'on transforme les contraintes de ce

problème en égalités

$$'A\,'u + z = 'c, \quad 'b\,'u \to \max. \tag{5}$$

Vu que les variables u ne sont pas obligées d'être positives, il n'y a aucune raison d'exiger la positivité des seconds membres.

4. Sommets et arêtes du polyèdre d'un problème canonique. Les sommets d'un ensemble polyédrique défini par le système d'inéquations linéaires sont déterminés par transformation en égalités exactes de n inéquations indépendantes quelconques du système. Le système (1) comprend m égalités

$$\begin{aligned} a_1^1 x^1 + \ldots + a_n^1 x^n &= b^1, \\ \ldots\ldots\ldots\ldots\ldots\ldots &\ldots\ldots \\ a_1^m x^1 + \ldots + a_n^m x^n &= b^m. \end{aligned} \tag{6}$$

Par conséquent, en chaque sommet, on doit remplacer par des égalités au moins $n - m$ inégalités $x^1 \geqslant 0, \ldots, x^n \geqslant 0$. Cela signifie que chaque sommet possède au moins $n - m$ coordonnées nulles.

Les sommets possédant plus de $n - m$ coordonnées nulles sont dits ***dégénérés***. Si un tel sommet existe, l'ensemble polyédrique et le problème sont dits ***dégénérés***.

Un exemple d'ensemble polyédrique dégénéré nous est fourni par le segment défini par le système d'inéquations

$$\begin{aligned} x^1 + x^2 + x^3 &= 1, \\ x^1 + \frac{1}{2} x^2 + 2x^3 &= 1, \\ \mathbf{x} &\geqslant \mathbf{0}. \end{aligned}$$

L'extrémité $N\,(1, 0, 0)$ de ce segment est un sommet dégénéré (fig. 56).

Un exemple d'ensemble polyédrique non dégénéré est donné par le système d'inéquations

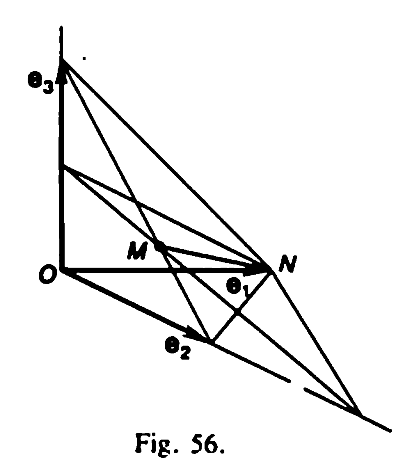


Fig. 56.

$$x^1 + x^2 + x^3 = 1,$$
$$\frac{1}{2}x^1 + 2x^2 + 2x^3 = 1,$$
$$x \geqslant 0,$$

qui définit le segment MN sur la figure 57. La comparaison de ces exemples montre que le dégénérescence est la propriété non pas de l'ensemble polyédrique mais du système d'inéquations qui le définit.

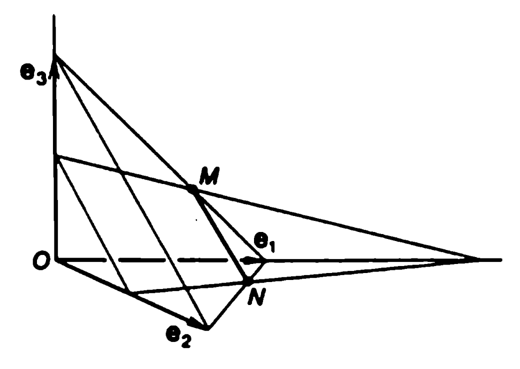


Fig. 57.

PROPOSITION 2. *Pour chaque sommet d'un ensemble polyédrique convexe défini par le système* (1) *il existe m colonnes linéairement indépendantes de la matrice A, dont toutes les colonnes correspondant aux coordonnées strictement positives du sommet considéré.*

DÉMONSTRATION. Supposons pour simplifier les notations que les $n - m$ dernières coordonnées du sommet donné sont nulles et considérons la matrice d'ordre n :

$$U = \left\| \frac{A}{O \mid E} \right\|,$$

où A est la matrice du système (1) et E la matrice unité d'ordre $n - m$. La matrice U est composée des lignes correspondant aux inéquations qui se transforment en équations pour le sommet considéré. Par hypothèse, ces inéquations sont linéairement indépendantes et, par suite, le déterminant de la matrice n'est pas nul. Il en découle qu'est différent de zéro le mineur d'ordre m de la matrice A, contenu dans ses m premières colonnes, ce qui achève la démonstration.

Les colonnes dont il s'agit dans la proposition 2 sont des colonnes du mineur principal de la matrice A. En programmation linéaire, les m colonnes linéairement indépendantes de la matrice A sont tout simplement appelées *base*. Les coordonnées du sommet associées aux colonnes d'une base sont dites *de base*.

La correspondance entre les bases et les sommets peut être décrite par les propositions suivantes :

a) A chaque sommet est associée au moins une base. Si le sommet est dégénéré (et seulement dans ce cas-là), on peut lui faire correspondre plus d'une base.

b) Un sommet n'est pas obligatoirement associé à une base. Si l'on suppose que les variables secondaires sont égales à zéro et que l'on obtienne les variables de base à partir du système, la matrice-colonne $\boldsymbol{x}^I$ à m éléments, composée de ces variables, est égale à $A_I^{-1}\boldsymbol{b}$, où A_I est une sous-matrice de la matrice A, composée des colonnes de la base. Donc, à une base correspond un sommet si

$$\boldsymbol{x}^I = A_I^{-1}\boldsymbol{b} \geqslant \boldsymbol{0},$$

et seulement dans ce cas-là. La base à laquelle correspond un sommet est dite *admissible*.

L'existence des problèmes dégénérés entraîne d'importantes difficultés en théorie de la méthode du simplexe. Dans les calculs suivant la méthode du simplexe, on doit donc prendre des mesures spéciales sans lesquelles l'algorithme peut s'avérer non efficace au cas d'un problème dégénéré.

On supposera dans la suite que le problème est non dégénéré, en reportant la discussion des problèmes dégénérés au point 8.

Une arête du polyèdre issue du sommet $\boldsymbol{x}_0$ peut être définie par l'équation paramétrique

$$\boldsymbol{x} = \boldsymbol{x}_0 + t\boldsymbol{l}, \tag{7}$$

où $\boldsymbol{l}$ est un vecteur directeur et t un paramètre positif. Si l'arête est un segment, $t \in [0, T]$, si elle est une demi-droite, $t \in [0, +\infty[$.

Vu que pour tous les points du polyèdre on a les égalités $A\boldsymbol{x} = \boldsymbol{b}$, le vecteur directeur de l'arête satisfait à la condition

$$A\boldsymbol{l} = \boldsymbol{0}. \tag{8}$$

Etant une face unidimensionnelle du polyèdre, l'arête se définit par un sous-système du système (1) dans lequel $n - 1$ inéquations sont remplacées par les équations. Il s'ensuit que tous les points de l'arête ont les mêmes $n - m - 1$ coordonnées nulles. Soit $I = \{i_1, \ldots, i_m\}$ l'ensemble des numéros de colonnes d'une base correspondant au sommet $\boldsymbol{x}_0$. Etant donné que le problème est supposé non dégénéré, cette base est définie parfaitement et $x_0^i > 0$ pour tous les $i \in I$.

Pour cette raison, chacune des coordonnées x^i, $i \in I$, plus encore une coordonnée dont le numéro sera noté k, des points de l'arête peuvent différer de zéro. Les autres coordonnées sont *a priori* nulles et, par suite,

$$l^j = 0, \quad j \notin I, \quad j \neq k. \tag{9}$$

Compte tenu de ces égalités dans la formule (8), il vient

$$A_I l^I + a_k l^k = \mathbf{0},$$

où A_I est la matrice composée de colonnes de la matrice A, dont les numéros appartiennent à I, et l^I la matrice-colonne à m éléments, composée des coordonnées l^i, $i \in I$. D'où

$$l^I = -l^k A_I^{-1} a_k. \tag{10}$$

Par substitution des valeurs (9) dans l'équation paramétrique (7) on obtient

$$\begin{aligned} x^I &= x_0^I + l^I t, \\ x^k &= l^k t, \\ x^j &= 0, \quad j \notin I, \quad j \neq k. \end{aligned} \tag{11}$$

On peut maintenant déterminer si l'arête est une demi-droite ou un segment et à quoi est égale dans ce dernier cas la valeur de t correspondant à la seconde extrémité du segment. Remarquons tout d'abord que

$$l^k > 0. \tag{12}$$

En effet, pour $l^k < 0$ on aurait $x^k < 0$ pour tous les $t > 0$, ce qui est impossible, et pour $l^k = 0$, selon (10), on aurait $l = \mathbf{0}$. A proprement parler, la condition $l^k > 0$ fixe la direction du vecteur l, défini par les formules (9) et (10) au facteur près, de manière que ce vecteur soit dirigé le long de l'arête vers le deuxième sommet (si l'arête est un segment) ou suivant la demi-droite (si l'arête est une demi-droite).

Pour $l^I \geqslant \mathbf{0}$, il est évident que la formule (11) définit pour tous les $t \geqslant 0$ une solution positive du système (6). Ceci revient à dire que l'arête considérée est une demi-droite.

Admettons que parmi les éléments de la matrice-colonne l^I il y a des éléments strictement négatifs. Notons I^- l'ensemble des i pour lesquels $l^i < 0$. Considérons la demi-droite définie par les formules (11) pour $t \geqslant 0$. Les coordonnées x^i, $i \in I^-$, du point courant de cette demi-droite changent de signe, chacune pour sa valeur de t. La coordonnée de numéro s pour lequel le quotient $-x_0^i/l^i$, $i \in I^-$, est minimal changera son signe la première. Elle s'annulera pour la valeur de t égale à

$$T = \min_{i \in I^-} \left[-\frac{x_0^i}{l^i} \right] = -\frac{x_0^s}{l^s}. \tag{13}$$

C'est justement la valeur du paramètre correspondant au deuxième sommet x_1 sur l'arête étudiée. Pour ce sommet on a $x_1^s = 0$ et $x_1^k > 0$. Aussi, la base associée au sommet x_1 s'obtient-elle à partir de la base du sommet x_0 par substitution de la colonne de numéro k à la colonne de numéro s. S'il

arrive que le minimum (13) est atteint pour au moins deux valeurs $i = s_1$, s_2, on peut associer deux bases au sommet x_1 qui est alors dégénéré. On suppose qu'il n'en est pas ainsi.

Remarquons que dans toutes les discussions précédentes le numéro k était arbitraire non compris dans I. Par suite, il part $n - m$ arêtes d'un sommet arbitraire non dégénéré.

On sait que pour une fonction φ de n variables la dérivée dans la direction définie par le vecteur l est la projection du gradient

$$\operatorname{grad} \varphi = \left\| \frac{\partial \varphi}{\partial x^1}, \ldots, \frac{\partial \varphi}{\partial x^n} \right\|$$

sur cette direction, c'est-à-dire

$$\sum_{i=1}^{n} \frac{\partial \varphi}{\partial x^i} l^i, \tag{14}$$

si le vecteur l est normé de manière que la somme des carrés de ses composantes soit égale à 1. La fonction $\varphi(x)$ décroît dans la direction de l si l'expression (14) est strictement négative. Le signe de cette expression est indépendant de la norme de l.

On considère une fonction linéaire $\varphi(x) = c_1 x^1 + \ldots + c_n x^n$ dont le gradient est constant :

$$\operatorname{grad} \varphi = \| c_1, \ldots, c_n \|$$

et sa projection sur l'arête en chaque point de cette dernière est la même. Selon (9) et (10) on a

$$\sum_{i=1}^{n} c_i l^i = c_I l^I + c_k l^k = l^k (c_k - c_I A_I^{-1} a_k),$$

où c_I et l^I sont les matrices-ligne et -colonne formées des composantes de c et l dont les numéros appartiennent à I. Vu que $l^k > 0$, le signe de la projection du gradient sur l'arête considérée coïncide avec le signe du nombre

$$\Delta_k = c_k - c_I A_I^{-1} a_k. \tag{15}$$

Il est naturel que les nombres Δ_k, $k \notin I$, jouent un rôle important dans la méthode du simplexe. On les appelle *estimations de substitution* correspondant à la base définie par l'ensemble de numéros I.

5. Etape de la méthode du simplexe. Il est maintenant clair quels sont les calculs à faire pour passer d'un sommet du polyèdre au sommet voisin et diminuer par là même la valeur de la fonction économique. Soit un som-

met de départ défini par l'ensemble I de numéros des colonnes de la base qui lui est associée.

Ce qu'on doit faire en premier lieu, c'est de calculer les estimations de substitution (15) pour tous les $k \notin I$. Si elles sont toutes positives, le sommet considéré est solution du problème, et il ne nous reste qu'à calculer ses coordonnées non nulles suivant la formule

$$x^I = A_I^{-1} b \tag{16}$$

et la valeur de la fonction

$$\varphi(x) = c_I x^I$$

pour obtenir la réponse.

Si parmi les estimations de substitution il y en a des strictement négatives, on recherche la maximale en module parmi ces dernières, en notant k son numéro. La matrice-colonne de numéro k sera introduite dans la base.

Ensuite, on recherche la matrice-colonne

$$l^I = -A_I^{-1} a_k \tag{10'}$$

(la norme du vecteur directeur de l'arête est telle que $l^k = 1$). Si tous les éléments de cette colonne sont positifs, le problème n'a pas de solution car la fonction économique n'est pas minorée.

Si parmi les éléments de la matrice-colonne l^I il y en a des strictement négatifs, on cherche le quotient minimal (13) et le numéro $s \in I$ pour lequel ce quotient est minimal. La colonne de numéro s doit être exclue de la base. On a ainsi défini un nouvel ensemble I' comprenant les numéros des colonnes de la base associée au nouveau sommet, et l'étape de la méthode du simplexe s'achève.

La principale difficulté dans ces calculs est la recherche de la matrice A_I^{-1} (si, évidemment, les dimensions de la matrice ne sont pas si grandes que, par exemple, un simple dénombrement entrepris pour choisir le rapport minimal (13) s'avère difficile). On peut remarquer qu'en réalité on a besoin non pas de la matrice A_I^{-1} mais des produits $c_I A_I^{-1}$, $A_I^{-1} b$, $-A_I^{-1} a_k$ qui peuvent être obtenus par résolution du système d'équations linéaires $yA_I = c_I$, $A_I x^I = b$ et $A_I l^I + a_k = \mathbf{0}$. Pour les résoudre, il suffit d'obtenir une seule fois la LU-décomposition de la matrice A_I. Une telle approche est mise en œuvre par exemple dans le programme contenu dans le livre de Wilkinson et Reinsch [43].

Les calculs de la méthode du simplexe peuvent se baser sur des transformations élémentaires effectuées sur les lignes de la matrice

$$T = \begin{Vmatrix} A & b \\ c & 0 \end{Vmatrix},$$

ou sous forme plus développée

$$T = \left\| \begin{matrix} a_1^1 & \dots & a_n^1 & b^1 \\ \multicolumn{4}{c}{\dots\dots\dots\dots\dots\dots} \\ a_1^m & \dots & a_n^m & b^m \\ c_1 & \dots & c_n & 0 \end{matrix} \right\| .$$

Cette matrice (aussi bien sous forme initiale qu'après transformations) est appelée *tableau du simplexe*. Si on connaît l'ensemble des indices de la base de départ I, on arrive, par des opérations élémentaires sur les lignes de la matrice T, à transformer ses colonnes de numéros $i \in I$ pour qu'elles deviennent les premières colonnes de la matrice unité d'ordre $m + 1$. On voit aisément que ces transformations sont équivalentes à la multiplication à gauche par la matrice

$$V = \left\| \begin{matrix} A_I^{-1} & 0 \\ -c_I A_I^{-1} & 1 \end{matrix} \right\| = \left\| \begin{matrix} E & 0 \\ -c_I & 1 \end{matrix} \right\| \cdot \left\| \begin{matrix} A_I^{-1} & 0 \\ 0 & 1 \end{matrix} \right\| .$$

Notons A_J la sous-matrice de la matrice A, constituée des colonnes n'appartenant pas à la base, et c_J la matrice-ligne des coefficients c_j pour lesquels $j \notin I$. Si pour plus de clarté on admet que les colonnes de base se trouvent aux premières places, on peut écrire

$$VT = \left\| \begin{matrix} A_I^{-1} & 0 \\ -c_I A_I^{-1} & 1 \end{matrix} \right\| \cdot \left\| \begin{matrix} A_I & A_J & b \\ c_I & c_J & 0 \end{matrix} \right\| = \left\| \begin{matrix} E & A_I^{-1} A_J & A_I^{-1} b \\ 0 & c_J - c_I A_I^{-1} A_J & -c_I A_I^{-1} b \end{matrix} \right\| .$$

Il s'avère évident que la matrice VT contient toute l'information nécessaire. Remarquons que l'élément droit inférieur de cette matrice ne diffère que par le signe de la valeur de la fonction économique dans le sommet considéré.

L'avantage important de ce procédé réside dans le fait qu'il n'est pas nécessaire de transformer la matrice T avec le passage au sommet suivant : il suffit de prendre VT pour matrice de départ et procéder à sa transformation. En effet, le système des contraintes du problème à matrice VT est équivalent à celui du problème de départ ; quant à la fonction économique, elle est obtenue de la fonction initiale φ par adjonction d'une expression identiquement nulle sur l'ensemble polyédrique défini par les contraintes.

La circonstance notée facilite de façon importante les calculs car une seule colonne de la base change en une étape et il faut appliquer à la matrice VT les seules transformations élémentaires qui correspondent à la colonne introduite dans la base.

Le défaut de la méthode fondée sur la transformation du tableau réside dans le fait qu'on est obligé au cours de la transformation de calculer tous

les produits $A_I^{-1}a_j$, $j \notin I$, tandis qu'on n'en a pas directement besoin. En effet, on n'a pas besoin de $A_I^{-1}A_J$. Sont exigés $(cA_I^{-1})A_J$ et $A_I^{-1}a_k$ pour la matrice-colonne a_k introduite dans la base.

Pour augmenter le rendement des calculs de la méthode du simplexe on peut procéder à la transformation de la matrice A_I^{-1} une fois trouvée en la matrice $A_{I'}^{-1}$ correspondant au sommet suivant. Ce n'est pas compliqué, vu que $A_{I'}$ se déduit de A_I par modification d'une seule colonne. On peut se servir ici de la forme d'élimination de la matrice inverse, utilisée non seulement dans la méthode du simplexe mais aussi dans différentes méthodes numériques d'algèbre linéaire.

6. Méthode d'élimination de la matrice inverse. Il nous faut nous rappeler comment on a obtenu la LU-décomposition par la méthode de Gauss, exposée au § 3 du ch. XIII. Soit B une matrice carrée régulière dont les lignes et colonnes sont ordonnées de manière que les mineurs principaux de B soient différents de zéro. A la page 402, on a introduit les matrices S_k telles que

$$B^{(k)} = S_k B^{(k-1)}, \quad k = 0, \ldots, n,$$

où $B^0 = B, B^1, \ldots, B^{(n)} = U$ est une suite de matrices déduites de B par la méthode de Gauss. Chaque matrice S_k ne diffère de la matrice unité que par une colonne et peut être écrite sous la forme

$$S_k = E + (\boldsymbol{\sigma}_k - \boldsymbol{e}_k)\,{}^t\boldsymbol{e}_k,$$

où $\boldsymbol{e}_k$ est une colonne de la matrice unité et les éléments de la colonne $\boldsymbol{\sigma}_k$ s'expriment au moyen d'éléments de $B^{(k-1)}$ suivant les formules

$$\begin{aligned} \sigma_k^i &= 0, \; i < k, \\ \sigma_k^k &= (b_{kk}^{(k-1)})^{-1}, \\ \sigma_k^i &= -b_{ik}^{(k-1)}/b_{kk}^{(k-1)}, \; i > k. \end{aligned}$$

Pour mémoriser S_k, il suffit d'écrire les éléments de la colonne $\boldsymbol{\sigma}_k$ de numéros $i \geqslant k$, vu que le produit de S_k par une matrice-colonne arbitraire $\boldsymbol{\xi}$ peut être trouvé facilement à l'aide de ces seuls éléments. En effet,

$$S_k\boldsymbol{\xi} = \boldsymbol{\xi} + (\boldsymbol{\sigma}_k - \boldsymbol{e}_k)\,{}^t\boldsymbol{e}_k\boldsymbol{\xi} = \boldsymbol{\xi} + (\boldsymbol{\sigma}_k - \boldsymbol{e}_k)\xi^k.$$

De façon analogue on peut obtenir le produit d'une matrice-ligne arbitraire par S_k.

On peut effectuer la suite d'opérations de substitution inverse dans le schéma de division unique en multipliant U par les matrices T_j, $j = n, \ldots$ $\ldots, 2$. Pour $i > j$, l'élément u_{ij} de la matrice U vaut $b_{ij}^{(i)}$. Aussi les $j - 1$ opérations élémentaires qui transforment la j-ième colonne de U en la j-ième colonne de la matrice unité sont-elles équivalentes à la multiplication

de U par la matrice

$$T_j = E + \tau_j({}^te_j),$$

où τ_j est une matrice-colonne d'éléments

$$\tau_j^i = \begin{cases} -b_{ij}^{(i)}, & i > j, \\ 0, & i < j. \end{cases}$$

(Rappelons que les éléments diagonaux de U sont égaux à l'unité.) De même que pour les matrices S_k, le produit $T_j\xi$ peut être calculé d'après les éléments de τ_j à l'aide d'un nombre peu élevé d'opérations arithmétiques :

$$T_j\xi = \xi + \xi^j\tau_j.$$

Après multiplication par $n - 1$ matrices la matrice U devient une matrice unité

$$T_2 \dots T_n U = E.$$

Cela veut dire que

$$T_2 \dots T_n S_n \dots S_1 B = E,$$

ou

$$B^{-1} = T_2 \dots T_n S_n \dots S_1.$$

Cette décomposition de B^{-1} en facteurs est appelée *forme d'élimination* de la matrice inverse. Cette décomposition contient n matrices triangulaires inférieures S_k et $n - 1$ matrices triangulaires supérieures T_j. Le lecteur peut trouver des renseignements plus détaillés sur la forme d'élimination et autres formes de représentation de la matrice inverse dans le livre de Twearson [37].

La forme d'élimination est particulièrement commode pour le traitement de grandes matrices creuses. Ces matrices se rencontrent souvent dans les problèmes de programmation linéaire.

Voyons comment se modifie la forme d'élimination de la matrice B^{-1} si on remplace dans B la j-ième colonne par la matrice-colonne $\bar{b}_j$. Désignons par $\bar{B}$ la matrice résultant de cette substitution. En effectuant les opérations élémentaires sur les lignes, on voit que chaque colonne se transforme indépendamment des autres. Aussi les matrices

$$H = T_{j+1} \dots T_n S_n \dots S_1 B$$

et

$$\bar{H} = T_{j+1} \dots T_n S_n \dots S_1 \bar{B}$$

ne diffèrent-elles l'une de l'autre que par la j-ième colonne. Soient α^i,

$i = 1, \ldots, n$, les éléments de la j-ième colonne de $\bar{H}$. Pour la transformer en colonne $\boldsymbol{e}_j$ il faut multiplier $\bar{H}$ par la matrice

$$\bar{T}_j = E + \bar{\boldsymbol{\tau}}_j({}'\boldsymbol{e}_j), \tag{17}$$

où $\bar{\boldsymbol{\tau}}_j$ est la matrice-colonne d'éléments

$$\bar{\tau}_j^i = \begin{cases} (\alpha^j)^{-1}, & i = j, \\ -\alpha^i/\alpha^j, & i \neq j. \end{cases} \tag{18}$$

(En décomposant det $\bar{H}$ suivant les dernières colonnes, on vérifie facilement que $\alpha^j \neq 0$.)

Les matrices $T_j H$ et $\bar{T}_j \bar{H}$ sont égales et, par suite, se réduisent à la matrice unité par multiplication par une même suite de matrices $T_2 \ldots T_{j-1}$. On voit que la forme d'élimination de $\bar{B}^{-1}$ se déduit de celle de B^{-1} par substitution de $\bar{T}_j$ à T_j, où $\bar{T}_j$ se définit par les formules (17) et (18), avec

$$\boldsymbol{\alpha} = T_{j+1} \ldots T_n S_n \ldots S_1 \bar{\boldsymbol{b}}_j.$$

Dans la littérature sur la programmation linéaire on trouve la description d'un grand nombre d'autres procédés de calcul de la matrice inverse lors du changement de base, mais on se limitera à cet exemple suffisamment simple et caractéristique.

7. Recherche de la base de départ. On connaît déjà beaucoup sur la méthode du simplexe mais on est incapable de l'appliquer car on ne peut trouver de sommet à partir duquel on puisse débuter. En effet, en choisissant une base arbitraire, on ne peut être sûr que les variables de base obtenues par la formule (16) seront positives.

Si la base de départ est inconnue d'avance, le problème de son choix est résolu grâce à l'augmentation du nombre de variables. On dit que la variable est *isolée* dans une équation si elle y figure avec un coefficient strictement positif et n'intervient dans aucune autre équation. On peut immédiatement indiquer la base de départ si chacune des équations du système (6) contient une variable qui y est isolée. Plus précisément, on peut considérer que la base de départ est formée des colonnes correspondant à ces variables.

Les variables isolées peuvent par exemple apparaître lorsqu'on transforme les inéquations en des équations. Ceci étant, il peut aussi surgir une variable qui intervient dans cette équation avec le coefficient -1.

Le procédé de construction de la base de départ consiste à introduire des variables isolées spéciales, dites *artificielles*, dans toutes les équations ne renfermant pas de variables isolées.

Si l'on introduit p variables artificielles, on passe de l'ensemble polyédrique convexe $\mathscr{A}$ de l'espace $\mathscr{R}_n$ à l'ensemble polyédrique convexe $\bar{\mathscr{A}}$ de

l'espace $\mathscr{R}_{n+p}$. L'ensemble $\bar{\mathscr{A}}$ contient obligatoirement au moins un point. En fait, il renferme même un sommet, car pour le système de contraintes qui le définit il existe une base admissible.

Pour trouver le sommet de l'ensemble polyédrique $\mathscr{A}$ (ou nous convaincre qu'il est vide), considérons le soi-disant *M-problème* : trouver sur l'ensemble polyédrique $\mathscr{A}$ le minimum de la fonction

$$\psi(x^1, \ldots, x^{n+p}) = \varphi(x^1, \ldots, x^n) + M(x^{n+1} + \ldots + x^{n+p}),$$

où le deuxième terme représente le produit de la somme des variables artificielles $x^{n+1}, \ldots, x^{n+p}$ par un facteur numérique noté traditionnellement M. Plus loin, on désignera les points de l'espace $\mathscr{R}_{n+p}$ par des lettres surmontées d'un trait :

$$\bar{x} = {}^t\| x^1, \ldots, x^n, x^{n+1}, \ldots, x^{n+p} \|,$$

et leurs projections sur $\mathscr{R}_n$ par les mêmes lettres sans trait :

$$x = {}^t\| x^1, \ldots, x^n \|.$$

Le point x sera identifié à

$${}^t\| x^1, \ldots, x^n, 0, \ldots, 0 \|.$$

Trois possibilités suivantes peuvent se présenter :

1. Le M-problème est résoluble pour un M et a pour solution $\bar{x}_0$ dont toutes les coordonnées x_0^{n+j} (correspondant aux variables artificielles) sont nulles. Dans ce cas, le problème initial est résoluble et x_0 est sa solution.

En effet, x_0 vérifie le système des contraintes du problème initial et, par suite, $x_0 \in \mathscr{A}$. De plus, les fonctions φ et ψ se confondent sur $\mathscr{A}$ et, par suite, pour tout x de $\mathscr{A}$

$$\varphi(x) \geqslant \min_{\bar{\mathscr{A}}} \psi. \tag{19}$$

Si pour x_0 on a dans ce cas une égalité, la valeur de $\varphi(x_0)$ est évidemment minimale sur $\mathscr{A}$.

2. Il existe un M suffisamment grand pour lequel le M-problème est résoluble, mais l'une quelconque de ses solutions $\bar{x}_0$ possède une coordonnée non nulle de numéro supérieur à n. Dans ce cas, le polyèdre $\mathscr{A}$ est un ensemble vide et le problème initial est irrésoluble.

En effet, si $\mathscr{A}$ n'est pas un ensemble vide, la résolubilité du M-problème entraîne, en vertu de (19), celle du problème initial et on a l'inégalité

$$\min_{\mathscr{A}} \varphi \geqslant \min_{\bar{\mathscr{A}}} \psi. \tag{20}$$

Notons ε la plus petite des coordonnées strictement positives x_0^{n+j} du

point $\bar{x}_0$. Il est évident que

$$\varphi(x_0) + M\varepsilon \leqslant \psi(\bar{x}_0) = \min_{\bar{\mathscr{A}}} \psi.$$

Prolongeons la fonction φ sur l'espace $\mathscr{R}_{n+p}$ suivant la formule $\varphi(\bar{x}) = \varphi(x)$ et désignons $\min\limits_{\bar{\mathscr{A}}} \varphi$ et $\min\limits_{\mathscr{A}} \varphi$ par $\varphi_{\bar{\mathscr{A}}}$ et $\varphi_{\mathscr{A}}$ respectivement. De la formule (20) il vient

$$\varphi_{\bar{\mathscr{A}}} + M\varepsilon \leqslant \varphi_{\mathscr{A}},$$

ou

$$M \leqslant \frac{\varphi_{\mathscr{A}} - \varphi_{\bar{\mathscr{A}}}}{\varepsilon}.$$

Ainsi donc, l'hypothèse que $\mathscr{A}$ n'est pas un ensemble vide aboutit à l'existence d'un majorant pour le nombre M. Ceci démontre l'assertion énoncée.

Il découle de l'analyse des cas 1 et 2 que pour un M suffisamment grand il ne peut exister deux solutions $\bar{x}_0$ et $\bar{x}_1$ du M-problème, telles que $x_0^{n+i} = 0$, $i = 1, \ldots, p$, mais existe un $x_1^{n+j} \neq 0$.

3. Le M-problème est irrésoluble pour des M suffisamment grands. Dans ce cas, le problème initial est aussi irrésoluble.

En effet, supposons que le problème initial est résoluble. Dans ce cas, le problème dual est aussi résoluble, c'est-à-dire qu'il existe une matrice-ligne u^0 à m éléments pour laquelle $u^0 A \geqslant c$, et la fonction ub prend la valeur maximale $u^0 b$.

Pour simplifier les notations on posera que les variables artificielles sont introduites dans toutes les contraintes du problème initial. Dans ce cas, le problème dual du M-problème est formulé de la façon suivante : chercher une matrice-ligne u à m éléments, telle que ub présente sa valeur maximale et soient vérifiées les contraintes $uA \leqslant c$ et $u_i \leqslant M$ pour tous les $i = 1, \ldots, m$. Si M est plus grand que la coordonnée maximale de u_0, alors u_0 vérifie le système des contraintes du M-problème dual, et ce système de contraintes est compatible. La fonction ub étant bornée sur l'ensemble des solutions du système $uA \leqslant c$, l'est encore après l'introduction de nouvelles contraintes. Il en découle que le M-problème dual est résoluble et par suite, est résoluble le M-problème. L'assertion est démontrée.

Les estimations obtenues pour savoir lequel des M peut être considéré « suffisamment grand » sont pratiquement inutiles. En utilisant cette méthode de recherche de la base de départ, on organise les calculs de manière que M soit supérieur à tout nombre qu'on lui compare. Alors en résolvant le M-problème, on obtient que les Δk de la formule (15) sont strictement positifs et supérieurs aux autres Δ_k pour $k > n$. Aussi les varia-

bles artificielles seront-elles exclues les premières du nombre des variables de base. Une fois qu'une variable artificielle sera exclue de la base, elle peut être tout à fait éliminée de l'examen.

S'il n'y a pas de raisons de douter de la compatibilité du système de contraintes du problème initial, on peut avec un certain risque (qui d'ailleurs n'est pas trop grand si l'on possède de l'expérience) fixer tout simplement pour M un nombre assez élevé. Si l'on obtient une solution du M-problème, dont les variables artificielles sont nulles, on peut dire que le choix de M a été bien fait.

8. Cyclage. On a appelé problème dégénéré de programmation linéaire un problème dans lequel à un sommet au moins de l'ensemble polyédrique de points admissibles il correspond non pas une seule base mais plusieurs. Le phénomène est lié à l'annulation des variables de base, ou, en langage géométrique, au fait que par le sommet considéré il passe plus de n hyperplans obtenus par transformation des contraintes du problème en égalités. Un sommet dégénéré peut être assimilé à deux (ou plusieurs) sommets confondus, ou à des sommets joints par une arête de longueur nulle.

Dans l'étude de la méthode du simplexe on a supposé que le problème est non dégénéré. Or, dans la pratique, les problèmes dégénérés se rencontrent assez souvent. Par exemple, si l'on débute par une base composée de colonnes correspondant aux variables isolées (voir point 7) et si dans la colonne des seconds membres il y a un élément nul, l'une des variables de base est égale à zéro, de sorte que le problème est dégénéré.

Ainsi, l'hypothèse que le problème n'est pas dégénéré est trop restrictive. Voyons les difficultés que peut entraîner la non-observation de cette restriction. La non-nullité des variables de base n'est pas en fait utilisée en elle-même. L'obstacle peut surgir du fait qu'à un même sommet sont associées plusieurs bases différentes. La projection du gradient de la fonction économique sur l'arête de longueur nulle n'est pas déterminée. Or, pour la recherche de la colonne qu'on doit éliminer de la base, on utilise non pas les projections mais les nombres Δ_j calculés par la formule (15) et qui sont toujours définis. Donc, si la nouvelle base est associée à l'ancien sommet, on ne s'en observera qu'à la fin de l'étape en remarquant que la valeur de la fonction économique n'a pas varié.

Ainsi, par l'algorithme de la méthode du simplexe on passe à une autre base, mais on peut rester au même sommet, de sorte qu'après l'étape de la méthode du simplexe la valeur de la fonction économique ne diminue pas. Or cela ne signifie pas encore que la recherche du minimum ne peut être menée jusqu'au bout. Mais il peut arriver qu'avec la règle adoptée de choix d'une colonne exclue de la base et d'une colonne introduite dans la base, on passe d'une base, associée au sommet considéré, à une autre base et qu'après une série d'étapes, on revienne à la base de départ. Ce phénomène est appelé *cyclage*.

Il faut avoir en vue que le cyclage n'est pas engendré par la méthode même, mais est en rapport avec un problème secondaire concernant la règle de choix des colonnes introduites dans la base et exclues de la base. Cette règle peut être améliorée de manière que le cyclage n'apparaisse pas. Le procédé suivant est peut-être le plus simple du point de vue de l'argumentation, mais il se peut qu'il entraîne une longue suite de calculs inutiles. S'il apparaît plusieurs variables, dont chacune peut être exclue de la base, cette variable est choisie de façon arbitraire. Vu que parmi les variables il y en a toujours celles dont l'exclusion de la base fournit une base associée à un autre sommet, quelques essais suffisent pour éliminer l'une d'elles.

Il existe aussi des procédés plus perfectionnés de lutte contre le cyclage, mais on n'y s'arrêtera pas. Le cyclage est un phénomène si rare qu'on est souvent obligé d'inventer spécialement des exemples de problèmes où il peut apparaître. Néanmoins, il faut toujours tenir compte de la possibilité de se heurter à un cyclage.

§ 5. Applications de la programmation linéaire

1. Problème de transport. De toutes les applications de la programmation linéaire on ne s'attachera ici qu'à quelques-unes. Aussi rigoureux que soit le choix de ces quelques applications, il semble qu'il faille choisir en premier lieu le soi-disant problème de transport. C'est un des premiers modèles d'optimisation linéaires, élaboré en 1939, avant que la programmation linéaire devienne une branche de mathématiques autonome. Pour le résoudre on a proposé une série de procédés devenus méthodes classiques de programmation linéaire.

Dans le cas le plus simple le *problème de transport* est formulé de la façon suivante. On considère m fournisseurs d'un certain bien et ses n consommateurs. Posons que les fournisseurs disposent des stocks de biens $a_1, \ldots, a_m$ et que les consommateurs doivent recevoir respectivement $b_1, \ldots, b_m$ unités de ce bien. Sont donnés les coûts de transport de l'unité du bien de chaque fournisseur à tout consommateur : c_{ij}, $i = 1, \ldots, m$, $j = 1, \ldots, n$. Il faut dans ces conditions établir le schéma des transports, c'est-à-dire indiquer pour tout couple ij la quantité de bien x_{ij} transportée du i-ième fournisseur au j-ième consommateur. De plus, il faut satisfaire les besoins de tous les consommateurs de la façon la plus économique et assurer la livraison de tous les biens des fournisseurs.

Il va de soi que ces exigences ne seront satisfaites que dans le cas où

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j .$$

Si cette condition est satisfaite, le problème de transport est dit ***équilibré***. Dans le cas général, le problème se réduit facilement au problème équilibré par introduction d'un fournisseur fictif ou d'un consommateur fictif.

Toutes les exigences imposées au schéma de transport, y incluse la sous-entendue positivité des nombres x_{ij}, peuvent être écrites sous la forme du problème suivant de programmation linéaire :

$$\sum_{i=1}^{m} x_{ij} = b_j, \quad \sum_{j=1}^{n} x_{ij} = a_i,$$

$$x_{ij} \geqslant 0, \quad \sum_{i,j} c_{ij}x_{ij} \to \min.$$

Le problème comprend mn variables et $m + n$ contraintes-égalités. La représentation du problème ne diffère de la forme canonique que par le fait que les contraintes sont linéairement dépendantes. On se convainc facilement de l'existence de la dépendance linéaire : la somme des égalités du premier groupe

$$\sum_{i,j} x_{ij} = \sum_{j} b_j$$

est la même que la somme des égalités du deuxième groupe

$$\sum_{i,j} x_{ij} = \sum_{i} a_i$$

(le problème est supposé équilibré). On n'élimine pas la dépendance linéaire des équations pour ne pas troubler la symétrie.

Notons T la matrice du système des contraintes. Chacune de ses colonnes contient deux unités et $m + n - 2$ zéros. Pour mieux nous représenter la structure de cette matrice, écrivons-la pour le cas de $m = 2, n = 3$. Les variables et les seconds membres des contraintes sont écrits pour identifier les colonnes et les lignes

$$T = \begin{array}{c} \begin{matrix} x_{11} & x_{12} & x_{13} & x_{21} & x_{22} & x_{23} \end{matrix} \\ \left\| \begin{matrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 \end{matrix} \right\| \begin{matrix} a_1 \\ a_2 \\ b_1 \\ b_2 \\ b_3 \end{matrix} \end{array}.$$

La matrice a de grandes (comparées à m et n) dimensions et une structure creuse. Aussi est-il naturel d'organiser les calculs de manière à ne pas

transformer la matrice. Grâce à sa structure spéciale cela est possible. Démontrons la propriété remarquable suivante de la matrice T.

PROPOSITION 1. *Toute sous-matrice régulière de la matrice T peut par permutation de lignes et colonnes être réduite à la forme triangulaire.*

Démontrons d'abord que dans toute sous-matrice régulière T' de T il existe obligatoirement une colonne contenant une unité. En effet, si toutes les lignes de la sous-matrice appartiennent au premier (ou second) groupe de lignes de la matrice T, chaque colonne contient une unité. Mais si la sous-matrice contient des lignes de différents groupes, additionnons les lignes du premier groupe et otons de la somme obtenue la somme des lignes du second groupe. Si chaque colonne renferme deux unités, ces unités sont situées dans les lignes de groupes différents et la différence obtenue est une ligne nulle.

Plaçons la colonne contenant une unité à la première place et transposons la ligne où se situe cette unité également à la première place. Maintenant on constate aisément que la sous-matrice T'' de T', située à l'intersection des lignes et colonnes de nouveaux numéros supérieurs à un, est également régulière. En reprenant les mêmes raisonnements que pour T', dégageons la matrice T''', etc.

Ainsi, en un nombre fini d'étapes, on transformera la matrice T' en matrice triangulaire supérieure.

Remarquons que chaque colonne de la matrice obtenue contient au-dessus de la diagonale une unité au plus, les autres éléments étant nuls. En outre, tous les éléments diagonaux sont aussi égaux à 1. On a donc la

PROPOSITION 2. *Tout mineur principal de la matrice T est égal à l'unité en valeur absolue.*

Déterminons maintenant le rang de la matrice T et en même temps indiquons le procédé de construction d'un sommet de l'ensemble polyédrique $\mathscr{T}$ des points admissibles du problème de transport.

Considérons une matrice C d'éléments c_{ij}. Soit $c_{\alpha\beta} = \min\limits_{i,j} c_{ij}$ et $a_\alpha < b_\beta$. En posant $x_{\alpha\beta} = a_\alpha$ et $x_{\alpha j} = 0$ pour tous les $j \neq \beta$, excluons de la discussion ultérieure la ligne de la matrice C de numéro α et remplaçons le nombre b_β par $b_\beta - a_\alpha$.

Si $a_\alpha > b_\beta$, on pose $x_{\alpha\beta} = b_\beta$ et $x_{i\beta} = 0$ pour $i \neq \alpha$. On remplace a_α par $a_\alpha - b_\beta$ et l'on exclut de la discussion la colonne de la matrice C de numéro β.

En cas d'égalité, on peut exclure de la même façon soit la ligne soit la colonne, au choix. Si l'on exclut par exemple une ligne, b_β est remplacé par zéro. Toutefois, si dans la matrice il y a une ligne et plusieurs colonnes, on

exclut la colonne, et s'il y a une colonne et plusieurs lignes, on exclut la ligne.

Une telle attribution des valeurs aux variables x_{ij} correspond à l'expédition, suivant l'itinéraire le moins couteux, d'autant de biens qu'il est possible, en satisfaisant de ce fait les besoins d'un consommateur ou en épuisant les stocks d'un fournisseur.

Après quoi on applique à la partie restante C' de la matrice C le même processus jusqu'à ce que soient exclues toutes les colonnes et les lignes. Il y a en tout $m + n$ colonnes et lignes et à chaque étape, sauf la dernière, est exclue une ligne ou une colonne. A la dernière étape, sont exclues et la ligne et la colonne. Ainsi, le processus comprend $m + n - 1$ étapes, si bien que la matrice construite $X = \| x_{ij} \|$ contient $m + n - 1$ éléments non nécessairement nuls.

Démontrons que les éléments de la matrice X ainsi construite vérifient le système des contraintes du problème de transport et que les variables dégagées sont des variables de base. Pour le démontrer, passons encore une fois en revue le processus de construction de la matrice X. Pour fixer les idées, posons qu'à la première étape on a exclu une ligne. La valeur de la variable $x_{\alpha_1\beta_1}$ dégagée à la première étape est parfaitement définie d'après a_{α_1} et b_{β_1}, à condition que $x_{\alpha_1 j} = 0$ pour tous les $j \neq \beta$. Cette valeur est telle que l'équation de numéro α_1 du premier groupe est vérifiée.

La matrice diminuée C' correspond au problème de transport équilibré dans lequel le nombre de consommateurs est diminué de 1, et $b'_{\beta_1} = b_{\beta_1} - a_{\alpha_1}$. La variable $x_{\alpha_2\beta_2}$ dégagée à la deuxième étape est aussi parfaitement définie d'après a'_{α_2} et b'_{β_2}, à condition que $x_{\alpha_2 j} = 0$ (ou respectivement $x_{i\beta_2} = 0$, si à la deuxième étape on a exclu une colonne). Vu que a'_{α_2} et b'_{β_2} sont définis d'après a_i et b_j, la valeur $x_{\alpha_2\beta_2}$ est parfaitement définie d'après a_i et b_j, et est satisfaite une contrainte correspondant à la matrice diminuée. Si $\beta_1 = \beta_2$, cette contrainte diffère de celle du problème initial, mais en tenant compte de la valeur $x_{\alpha_1\beta_1}$ et de la variation de b_{β_1}, on constate qu'est aussi vérifiée la contrainte du problème initial.

En raisonnant de la sorte, on obtient en fin de compte une matrice réduite à une ligne et à une colonne et deux nombres égaux a et b. On suppose que la dernière variable est égale à a, de sorte que les deux dernières équations qui restent se voient vérifiées.

Ainsi, toutes les équations sont vérifiées, et les variables dégagées sont définies parfaitement comme polynômes linéaires par rapport à $a_1, \ldots, a_m$ et $b_1, \ldots, b_n$, à condition que les autres variables soient nulles. Cela veut dire que les variables dégagées sont de base. Il en découle que

$$\operatorname{Rg} T = m + n - 1.$$

Remarquons que l'exigence de choisir à chaque étape l'élément minimal de la matrice C n'a été nullement utilisée. On aurait pu choisir chaque fois le premier élément venu (par exemple, l'élément supérieur gauche). Le choix de l'élément minimal fournit un sommet du polyèdre plus proche de la solution du problème. D'autre part, il existe des méthodes dont la réalisation est plus compliquée mais qui permettent d'obtenir des solutions initiales meilleures.

On constate que la construction de ıa base de départ dans le problème de transport ne présente pas de difficultés.

La structure décrite du mineur principal de la matrice T montre que les calculs suivant les formules (10), (15) et (16) du § 4 s'effectuent de façon très simple : pour résoudre les systèmes correspondants d'équations linéaires il faut faire assez peu d'additions et de soustractions et on n'a absolument pas besoin de multiplications et de divisions. Par conséquent, toutes les difficultés rencontrées dans l'application de la méthode du simplexe au problème de transport se réduisent au choix des variables qu'on introduit dans la base et qu'on exclut de la base. Cela permet de créer pour le problème de transport une modification de la méthode du simplexe d'un assez haut rendement. Cette modification (dénommée *méthode des potentiels*) a été construite avant la méthode du simplexe générale. On n'étudiera pas la méthode des potentiels en lui préférant l'étude d'autres problèmes liés au problème de transport.

2. Problème du flot maximal. Parmi les problèmes proches du problème de transport, étudions le *problème du flot maximal dans un réseau de transport*. On envisage dans ce problème un flot de marchandises ou d'autre chose (par exemple, de liquide) suivant un réseau de communications reliant certains sommets ou nœuds du réseau. Les arcs constituant le réseau sont supposés orientés. Si le mouvement entre deux nœuds est possible dans deux directions, on considérera que les nœuds sont réunis par deux arcs orientés de façon opposée.

A chaque arc constituant le réseau est associée une *capacité*, c'est-à-dire un nombre qui fixe la valeur maximale de débit dans cet arc. En certains nœuds se vérifient les équations d'équilibre : l'ensemble de flots entrant dans le nœud est égal à l'ensemble de flots sortant de ce nœud. Mais il existe des nœuds où ces ensembles de flots sont différents. Ces nœuds constituent les *origines* (*entrées*) ou *destinations* (*sorties*) suivant le signe du débit somme. Il est toujours possible d'introduire des arcs complémentaires de capacités correspondantes pour réunir toutes les origines en une seule et toutes les destinations en une seule.

Au modèle de réseau décrit on peut associer plusieurs problèmes d'optimisation. Le problème du flot maximal a pour objectif de rechercher la

valeur maximal du débit de l'origine vers la destination, qui soit compatible avec les capacités données.

Posons que le réseau comprenne n nœuds outre l'origine et la destination. Attribuons à l'origine le numéro 0 et à la destination le numéro $n + 1$. Le réseau peut être défini par une matrice D d'ordre $n + 1$. Un élément d_{ij}, $i = 0, \ldots, n$; $j = 1, \ldots, n + 1$, de cette matrice est égal à la capacité de l'arc si les nœuds i et j sont réunis par un arc et à zéro dans les autres cas. En particulier, les éléments diagonaux d_{ii} sont nuls.

Il faut remarquer que pour la plupart des réseaux réels la matrice D est creuse et, par suite, rien ne sert de l'écrire entièrement. Les données sur le réseau peuvent être aussi mémorisées sous d'autres formes. Toutefois, pour des raisonnements théoriques il est commode de se servir de la représentation matricielle.

Notons x_{ij} la valeur du débit dans l'arc du i-ième nœud vers le j-ième. Pour simplifier l'écriture des équations posons que les valeurs des débits sont définies pour tous les couples i, j, $i = 0, \ldots, n$; $j = 1, \ldots, n + 1$, et qu'on associe aux arcs manquants les capacités et flots nuls. Dans ce cas, les contraintes imposées aux flots par les capacités des arcs prennent la forme de

$$0 \leqslant x_{ij} \leqslant d_{ij} \tag{1}$$

pour tous les couples i, j.

L'équation d'équilibre en j-ième nœud, $j = 1, \ldots, n$, peut être écrite ainsi :

$$\sum_{i=0}^{n} x_{ij} - \sum_{k=1}^{n+1} x_{jk} = 0. \tag{2}$$

Il ressort de l'équation d'équilibre que la somme des débits sortant de l'origine doit être égale à la somme des débits aboutissant à la déstination :

$$\sum_{j=1}^{n+1} x_{0j} = \sum_{j=0}^{n} x_{j,\,n+1}. \tag{3}$$

La dernière équation peut aussi être prise pour équation d'équilibre si l'on suppose que la destination et l'origine sont réunies par un arc de grande capacité qui renvoit à l'origine tout ce qui a été écoulé à travers le réseau et réunit ainsi l'origine et la destination en un seul point.

Par valeur du flot dans le réseau on entendra le premier membre de l'égalité (3). Cette somme doit être rendue maximale à condition que soient vérifiées les contraintes (1), (2).

Pour réduire le problème à la forme canonique, introduisons des variables supplémentaires y_{ij}, $i = 0, \ldots, n$; $j = 1, \ldots, n + 1$, telles que les con-

traintes imposées aux capacités s'écrivent sous la forme d'égalités

$$x_{ij} + y_{ij} = d_{ij}. \tag{4}$$

Les variables supplémentaires y_{ij} sont positives.

La matrice S du système des contraintes (2), (4) possède deux groupes de lignes : les lignes du premier groupe sont constituées des coefficients des équations (2) et les lignes du second, des coefficients de (4). Chaque variable x_{ij} figure avec un coefficient non nul dans deux équations du premier groupe : dans l'une d'elle, elle intervient dans la première somme, et dans l'autre, dans la seconde. A chaque variable x_{ij} correspond une équation du second groupe, laquelle contient la variable y_{ij} avec les mêmes valeurs des indices.

Ainsi, la colonne de la matrice S correspondant à la variable x_{ij} contient trois éléments non nuls, à savoir, $+1$ et -1 dans les lignes du premier groupe et $+1$ dans les lignes du second groupe. Les colonnes correspondant aux variables y_{ij} contiennent une unité. Par analogie à la proposition 1 on peut démontrer la

PROPOSITION 3. *Chaque sous-matrice carrée régulière extraite de la matrice S du système de contraintes* (2), (4) *peut par permutation de lignes et colonnes être réduite à la forme triangulaire.*

Il suffit de démontrer que chaque sous-matrice régulière possède une colonne avec un élément non nul. Démontrons-le.

Si la sous-matrice renferme une colonne correspondant à y_{ij}, la démonstration n'est pas nécessaire : l'unique unité de cette colonne doit être contenue dans la sous-matrice. On admettra donc que ces colonnes ne figurent pas dans la sous-matrice considérée.

Si la sous-matrice ne contient que des lignes du second groupe, chaque colonne renferme une seule unité. Mais si les lignes appartiennent aux deux groupes, le déterminant de la sous-matrice se décompose en produit du mineur contenant toutes les unités des lignes du second groupe par son mineur associé, situé entièrement sur les lignes du premier groupe. Le problème se ramène ainsi à l'étude des sous-matrices contenues entièrement dans les lignes du premier groupe. A ces sous-matrices on applique la démonstration par absurde, absolument analogue à celle qui a été faite pour la proposition 1.

Il est évidemment inutile d'introduire dans la matrice S des colonnes correspondant à des variables égales à zéro. La représentation de la matrice est fonction du mode de définition choisi pour le réseau. Il nous faut toutefois noter que la proposition 3 reste encore vraie si on exclut de la matrice S des colonnes quelconques.

Posons que la matrice S contient des colonnes correspondant à N cou-

ples différents de valeurs d'indices (i, j). L'ensemble des variables de base d'une base quelconque contient obligatoirement pour chaque couple (i, j) soit x_{ij}, soit y_{ij}, soit ces deux variables, vu que l'équation $x_{ij} + y_{ij} = d_{ij}$, comme toute équation, doit contenir au moins une variable de base. En conséquence, les couples (i, j) se répartissent en trois classes :

G_1 : les deux variables x_{ij} et y_{ij} sont de base ;
G_2 : la variable x_{ij} est de base, y_{ij} ne l'est pas ;
G_3 : la variable y_{ij} est de base, x_{ij} ne l'est pas.

Le nombre de couples de la première classe peut être calculé. Si ces couples sont au nombre de k, les autres sont au nombre de $N - k$ et le nombre total des variables de base est $2k + N - k$. Or le nombre de variables de base est égal à celui d'équations indépendantes, c'est-à-dire à $n + N$. Il en découle que $k = n$.

Considérons la sous-matrice composée des colonnes de base et permutons les lignes et les colonnes de manière que les colonnes correspondant aux variables de base y_{ij} soient les dernières et les lignes contenant des unités dans ces colonnes soient de même les dernières. Il s'ensuit que la sous-matrice donnée se décompose en blocs :

$$\begin{Vmatrix} S_1 & O \\ S_2 & S_3 \end{Vmatrix}.$$

Il est évident que $\det S_1 \neq 0$ et que S_1 doit contenir la colonne renfermant un seul élément non nul. Or S_1 contient toutes les lignes du premier groupe et, par suite, cette colonne ne peut être que la colonne correspondant à la variable x_{0i} où a $x_{j,\,n+1}$. De plus, la deuxième unité de cette colonne n'intervient pas dans S_1, donç le couple $(0, i)$ ou $(j, n+1)$ se rapporte aux couples de la première classe G_1.

Formulons le dual du problème du flot maximal. Pour le problème de maximisation d'une fonction, étant donné la forme canonique des contraintes, le problème dual peut être formulé de la façon analogue au problème (4) du § 4 :

$$yA \geqslant c, \quad yb \to \min.$$

Dans le cas considéré, le problème dual contient n variables u_k correspondant aux contraintes (2), et N variables v_{ij} correspondant aux contraintes (4). On peut admettre que u_k correspondent aux nœuds du réseau et v_{ij} à ses arcs. Le système des contraintes du problème dual est de la forme

$$\begin{aligned} u_k + v_{0k} &\geqslant 1, \\ -u_l + v_{l,\,n+1} &\geqslant 0, \\ -u_k + u_l + v_{kl} &\geqslant 0, \\ v_{ij} &\geqslant 0. \end{aligned}$$

Ici et plus loin les indices prennent les valeurs suivantes : $i = 0, \ldots, n$; $j = 1, \ldots, n + 1$; $k, l = 1, \ldots, n$ qui se combinent de la façon déterminée, car le système contient des contraintes pour les seuls couples d'indices qui correspondent à N variables x_{ij} du problème primal.

Posons que u_k^*, v_{ij}^* est la solution du problème dual et x_{ij}^*, y_{ij}^* la solution du problème primal. Supposons que le problème est non dégénéré et considérons la base du problème primal correspondant à la solution. On dira que l'arc du réseau et la variable v_{ij} appartiennent à la classe G_1, G_2 ou G_3 si le couple d'indices correspondant (i, j) se rapporte à cette classe.

Pour les couples d'indices de la classe G_1 on a $x_{ij}^* > 0$, $y_{ij}^* > 0$. Cela veut dire que l'arc est chargé par un débit qui ne dépasse pas sa capacité. Selon la proposition 3 du § 3, u_k^* et v_{ij}^* vérifient les relations :

$$\begin{aligned} u_k^* &= 1 && \text{si} && 0, k \in G_1, \\ u_k^* &= 0 && \text{si} && k, n+1 \in G_1, \\ u_k^* &= u_l^* && \text{si} && k, l \in G_1, \\ v_{ij}^* &= 0 && \text{si} && l, j \in G_1. \end{aligned}$$

Pour u_k on a ici n relations indépendantes, autant que le nombre de couples dans G_1. Faïsons correspondre à l'origine $u_0^* = 1$ et à la destination $u_{n+1}^* = 0$. On peut alors en déduire que $u_i^* = 1$ pour tous les nœuds du réseau reliés à l'origine par des arcs non surchargés à la limite, et $u_j^* = 0$ pour les nœuds reliés à la destination par des arcs non chargés. Ceci étant, chaque variable u_k possède la valeur u_k^* égale à zéro ou à l'unité.

Pour les couples d'indices de la classe G_2 on a $x_{ij}^* > 0$, $y_{ij}^* = 0$. A cette classe appartiennent les arcs surchargés à la limite. On a pour ces derniers

$$\begin{aligned} v_{ij}^* > 0, &\quad v_{kl}^* = u_k^* - v_l^*, \\ u_k^* + v_{0k}^* = 1 &\quad \text{si} \qquad 0, k \in G_2, \\ u_k^* - v_{k,\,n+1}^* = 0 &\quad \text{si} \qquad k, n+1 \in G_2. \end{aligned}$$

Il s'ensuit que pour $i, j \in G_2$ on a $u_i^* = 1$ et $u_j^* = 0$. Cela veut dire que les arcs surchargés à la limite relient les nœuds pour lesquels $u_i^* = 1$ avec les nœuds pour lesquels $u_j^* = 0$.

Pour les couples d'indices de la classe G_3 on a $x_{ij}^* = 0$, $y_{ij}^* > 0$. Ce sont des arcs non chargés qu'on aurait pu exclure du réseau sans modifier le flot maximal. Il vient

$$v_{ij}^* = 0 \quad \text{si} \quad i, j \in G_3$$

et

$$u_k^* < u_l^* \quad \text{si} \quad k, l \in G_3,$$

ce qui signifie que les arcs libres relient les sommets pour lesquels $u_k^* = 0$ à ceux pour lesquels $u_l^* = 1$.

La fonction économique du problème dual est

$$\sum_{i,j} d_{ij} v_{ij}.$$

En vertu de ce qui a été dit plus haut, sa valeur minimale vaut

$$\sum_{i,j} d_{ij} v_{ij}^* = \sum_{i,j \in G_2} d_{ij}. \tag{5}$$

Dans la dernière somme, la sommation ne s'étend qu'aux couples de la classe G_2 car pour eux seuls v_{ij}^* est différent de zéro.

Voyons l'interprétation de ce résultat en termes du réseau. La somme (5) est la somme des capacités d'un ensemble d'arcs. Introduisons la définition suivante.

Définition. Admettons que l'ensemble des nœuds du réseau est subdivisé en deux sous-ensembles disjoints $\mathscr{P}_1$ et $\mathscr{P}_2$, et que l'origine appartient à $\mathscr{P}_1$ et la destination à $\mathscr{P}_2$. L'ensemble de tous les arcs dont l'origine appartient à $\mathscr{P}_1$ et l'extrémité à $\mathscr{P}_2$ est appelé *coupe* du réseau. La somme des capacités de tous les arcs de la coupe sera appelée *capacité* de cette coupe.

L'introduction du terme « coupe » s'appuie sur la propriété suivante. Considérons un *chemin* allant de l'origine à la destination. C'est une suite de nœuds $u_0, u_{i_1}, \ldots, u_{i_s}, u_{n+1}$ dont chacun est relié au suivant par un arc. La suite commence à l'origine u_0 et aboutit à la destination u_{n+1}. Quelle que soit la coupe du réseau, chaque chemin allant de l'origine à la destination contient obligatoirement un arc de cette coupe. En effet, suivons le chemin en notant auquel des sous-ensembles $\mathscr{P}_1$ ou $\mathscr{P}_2$ se rapportent les nœuds traversés. On commence par un nœud de $\mathscr{P}_1$ et l'on doit aboutir à un nœud de $\mathscr{P}_2$. Aussi doit-on passer quelque part de $\mathscr{P}_1$ à $\mathscr{P}_2$.

Si l'on rapporte à $\mathscr{P}_1$ les nœuds pour lesquels $u_i^* = 1$ et à $\mathscr{P}_2$ ceux pour lesquels $u_i^* = 0$, cette répartition définira la coupe composée des arcs de la classe G_2. La somme (5) est la capacité de cette coupe.

On dira que la coupe est *de capacité minimale* si sa capacité est inférieure à celle de toute autre coupe. On voit aisément que la valeur du flot maximal dans le réseau ne peut dépasser la capacité minimale de la coupe. En effet, avant d'atteindre la destination, le flot maximal doit passer dans l'ensemble $\mathscr{P}_2$ définissant la coupe de capacité minimale.

Maintenant le théorème de dualité appliqué au problème du flot maximal peut être formulé de la façon suivante.

Théorème 1. *La valeur du flot maximal dans un réseau est égale à la capacité minimale de la coupe de ce réseau.*

3. Programmation linéaire en nombres entiers. Une série d'importants modèles pratiques conduit aux problèmes de programmation linéaire avec une condition supplémentaire imposée à toutes (ou certaines) variables d'être entières.

Supposons par exemple qu'aux points A_i, $i = 1, \ldots, m$, on a organisé la production d'un bien en quantités a_i dont le coût est f_i. Les consommateurs aux points B_j ont des demandes b_j, $j = 1, \ldots, n$, qu'on doit satisfaire de la façon la plus économique en tenant compte des coûts c_{ij} de transport de A_i à B_j.

On aboutit ainsi au problème de programmation linéaire à fonction économique

$$\sum_{i,j} c_{ij} x_{ij} + \sum_i f_i y_i,$$

où $y_i = 0$ ou 1 suivant que la production est organisée au point A_i ou ailleurs. Ce problème appelé *problème* simple *de distribution de la production* est du même genre que celui de transport, étant toutefois plus compliqué à cause justement de la condition imposée à y_i d'être entier.

On aurait pu croire que la solution d'un problème de programmation linéaire en nombres entiers peut être obtenue comme une solution convenablement arrondie du problème de programmation linéaire ordinaire si on rejette les conditions l'obligeant à être entière. Mais en réalité une telle opinion est fausse, comme le montre la figure 58. Sur cette dernière, la partie hachurée représente l'ensemble polyédrique des points admissibles du problème de programmation linéaire à deux variables, et le vecteur c indique la direction du gradient de la fonction économique. La solution X du problème de programmation linéaire discrète diffère sensiblement de la solution Y du problème continu.

Les problèmes de programmation linéaire discrète, considérés sous l'angle des questions soulevées par leur résolution et des méthodes permettant de les surmonter, sont beaucoup plus proches des problèmes généraux de programmation discrète que des problèmes de programmation linéaire.

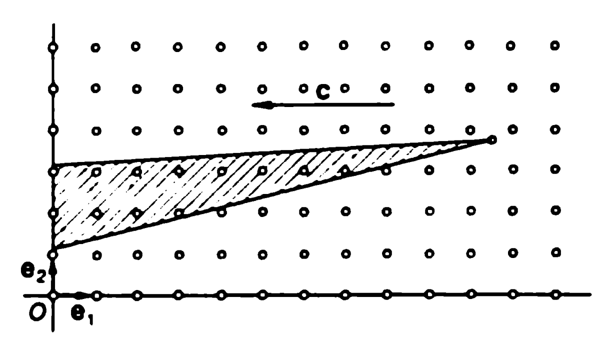


Fig. 58.

Il existe toutefois quelques problèmes discrets dont la solution peut être obtenue à partir du problème linéaire correspondant. Ils sont liés aux problèmes de transport et de réseau déjà étudiés, dont l'un sera maintenant abordé.

Le *problème d'affectation* se présente ainsi. Soient n postes et autant de candidats à ces postes, chaque candidat satisfaisant de façon différente aux exigences imposées pour occuper chacun de ces postes. (Il se peut, évidemment, que ce soient non pas des postes et des candidats devant les occuper, mais disons, des entreprises de construction et des machines qui y sont utilisées, etc.) Admettons que le profit attendu de l'affectation du i-ième candidat au j-ième poste est c_{ij}, $i, j = 1, \ldots, n$. Il est nécessaire d'établir une telle liste d'affectations des candidats aux postes qui rende maximal le profit total.

De par sa nature, le problème n'est aucunement lié à la programmation linéaire, et la première idée venue est de faire le dénombrement de $n!$ listes d'affectations possibles. Cependant, si n est assez élevé, ce procédé est peu attirant. En programmation discrète, on étudie divers procédés de dénombrement des variantes qui permettent de ne pas les passer toutes en revue et d'exclure par groupes entiers celles qui sont non conformes au problème. Essayons toutefois de poser le problème de programmation linéaire.

Introduisons les variables x_{ij} et admettons que $x_{ij} = 1$ si le i-ième candidat est affecté au j-ième poste, et $x_{ij} = 0$ s'il n'y est pas affecté. Dans ce cas, le profit attendu de l'adoption de la liste définie par la matrice $X = \| x_{ij} \|$ est égal à

$$\sum_{i,j} c_{ij} x_{ij}. \tag{6}$$

Cette somme doit être rendue maximale à la condition que X est une matrice de permutation, c'est-à-dire que chacune de ses lignes et chacune de ses colonnes contiennent une unité et une seule, les autres éléments étant nuls.

Si X est la matrice de permutation, on a

$$x_{ij} \geqslant 0, \quad \sum_{i=1}^{n} x_{ij} = 1, \quad \sum_{j=1}^{n} x_{ij} = 1. \tag{7}$$

Toutefois, ces conditions sont évidemment insuffisantes et ne garantissent pas que toutes les variables x_{ij} soient égales à 1 ou à 0.

Ainsi, la solution du problème de maximisation de la somme (6) avec conditions (7) fournit la solution du problème d'affectation si ses composantes x_{ij} sont nulles ou égales à l'unité, et ne résout pas le problème dans le cas contraire.

Remarquons que le problème formulé diffère du problème de transport par le seul fait qu'on doit maximiser la fonction donnée et non pas la minimiser. Cette différence n'a aucun rapport aux propriétés de la matrice du système de contraintes, si bien que cette matrice vérifie la proposition 2. Il ressort aussitôt de la proposition 2 la *propriété des solutions entières* du problème de transport : si les nombres a_i et b_j sont des entiers, tous les sommets du polyèdre $\mathscr{T}$ des points admissibles du problème de transport n'ont que des coordonnées entières.

Notons qu'une propriété analogue des solutions entières du problème du flot maximal découle de la proposition 3 si les capacités sont des nombres entiers.

Ainsi donc, on peut être sûr que pour un problème avec contraintes (7) le maximum de la fonction (6) est atteint au point de coordonnées entières. Si la solution n'est pas unique, il y aura aussi parmi les solutions des non entières, mais la solution entière existe obligatoirement. Les variables entières x_{ij} sont positives et les sommes dans (7) sont égales à l'unité, si bien que chaque x_{ij} est égale soit à zéro, soit à l'unité.

On voit que la résolution du problème discret s'est ramenée à celle d'un problème de programmation linéaire beaucoup plus simple et ceci grâce à la propriété de la solution entière du problème de transport.

Un exposé détaillé de la théorie des problèmes de réseaux et de leur application aux problèmes de programmation discrète peut être trouvé dans le livre de Hu [15].

4. Jeux matriciels. Les jeux matriciels forment une classe importante de modèles étroitement liés à la programmation linéaire.

Du point de vue mathématique on entend par *jeu* un processus dans lequel plusieurs personnes prennent des décisions. A la fin du processus, chaque participant aboutit à un gain (non nécessairement positif) qui dépend de ses décisions et de celles des autres joueurs. Outre les jeux au sens courant du terme, s'y rapportent les modèles de situations étudiés en stratégie et tactique, les modèles d'opérations boursières et nombre d'autres.

Dans les jeux matriciels qu'on étudiera, participent deux personnes dont les intérêts sont opposés car la somme de leurs gains est nulle. Etant donné cette hypothèse, le jeu est dit *antagoniste* ou *à somme nulle*. En réalité, les intérêts des joueurs peuvent aussi être opposés pour une somme de gains non nulle, mais ce cas se réduit facilement à celui de la somme nulle.

La prise de décision est supposée unique, c'est-à-dire que le résultat du jeu est déterminé après que chacun des joueurs ait pris une fois la décision. Les décisions prises par les joueurs sont appelées *stratégies*. Dans les jeux tels que les échecs, les adversaires font plusieurs coups en prenant chaque fois une décision. Dans ce cas, le jeu est de la classe des jeux *dynamiques*.

Mais on peut imaginer une situation dans laquelle un joueur, avant de commencer le jeu, décide de tous les coups qu'il fera (il va de soi que pour les échecs, comme pour tous les jeux réels, ce n'est qu'une possibilité théorique). Chacune de ces décisions constituera la stratégie du jeu à décision unique, si bien que le jeu reprend ainsi la *forme normale.*

Pour les jeux de forme normale, il est supposé qu'aucun des joueurs ne possède d'information certaine sur la stratégie choisie par l'autre, autrement dit les décisions sont prises par les joueurs de façon indépendante. On entend par information incertaine l'information que le joueur peut extraire des résultats précédents s'il a déjà joué avec cet adversaire. C'est ainsi que si l'adversaire utilisait toujours la stratégie α, on peut supposer qu'il la suivra encore.

Les jeux matriciels appartiennent à la classe des jeux *finis*. Cela suppose que le nombre de stratégies de chaque joueur est fini.

Toutes les hypothèses sont ainsi faites. Tout jeu antagoniste fini de forme normale est appelé *jeu matriciel.* Ce jeu peut être décrit par une matrice dont les lignes correspondent aux stratégies du premier joueur et les colonnes à celles du second. Un élément a_{ij} de la matrice est le gain obtenu par le premier joueur s'il choisit une stratégie de numéro i, tandis que celle de son adversaire est de numéro j. Si les joueurs changent de rôle, les lignes dans la matrice deviennent colonnes et *vice versa,* et les éléments changent de signe. Ainsi, la matrice A se remplace par la matrice $-{}^tA$. Un jeu *symétrique* par exemple, dans lequel les deux joueurs sont dans une même situation doit avoir une matrice symétrique gauche.

Il va de soi que pour définir un jeu il suffit de donner la matrice du premier joueur. C'est ainsi qu'on agira, bien qu'on introduise de la sorte une certaine asymétrie dans les notations. On étudiera un jeu défini par la matrice A d'éléments a_{ij}, $i = 1, \ldots, m$; $j = 1, \ldots, n$. Le premier joueur a m stratégies, le second, n. On peut donc considérer que le jeu consiste en ce que le premier joueur choisit une ligne de la matrice, tandis que le second, indépendamment du premier, choisit une colonne. Si à l'intersection de la ligne et de la colonne se trouve l'élément a_{ij}, le premier joueur obtient a_{ij} et le second, $-a_{ij}$.

Ainsi, le premier joueur cherche à obtenir l'élément maximal de la matrice et le second, l'élément minimal. En conséquence, on dit parfois que le premier est *joueur du maximum* et le second, *joueur du minimum.*

5. Gains garantis. Admettons que le premier joueur est une personne très prudente, qui ne veut pas risquer et cherche à obtenir, si possible, un gain garanti, non pas le plus grand gain mais qui ne dépend pas des décisions du second joueur. Le premier joueur peut raisonner de la sorte : « Dans le pire des cas, j'obtiendrai le plus petit élément de la ligne choisie. Il faut donc choisir la ligne dans laquelle le plus petit élément est maxi-

mal ». Ainsi, soit $i = 1, \ldots, m$; $j = 1, \ldots, n$ et

$$a_{i_0 j_0} = \max_i \min_j a_{ij}.$$

En choisissant la stratégie i_0, le premier joueur ne peut obtenir moins que $a_{i_0 j_0}$, quoi que le second joueur n'ait choisi. Pour d'autres stratégies il peut obtenir moins ou plus que $a_{i_0 j_0}$, ce qui dépend déjà des décisions du second joueur. Mais si le second agit de façon prudente, le premier ne peut obtenir plus d'une certaine limite qu'on va établir.

Le nombre $\max_i \min_j a_{ij}$ est appelé *gain garanti* du premier joueur. Le nombre $-\max_i \min_j a_{ij}$ est la limite supérieure du gain du second joueur qu'il ne peut dépasser en cas de comportement prudent du premier joueur.

Faisons un raisonnement analogue pour le second joueur. Pour obtenir un gain garanti il doit choisir la colonne dans laquelle le plus grand élément soit minimal, c'est-à-dire la colonne de numéro j_1 pour lequel

$$a_{i_1 j_1} = \min_j \max_i a_{ij}.$$

Ayant choisi la stratégie j_1, le second joueur obtient un gain garanti égal à $-\min_j \max_i a_{ij}$. Simultanément, on obtient que le nombre $\min_j \max_i a_{ij}$ est la limite supérieure du gain du premier joueur qu'il ne peut dépasser en cas de comportement prudent du second joueur.

Etablissons une relation entre les gains garantis des joueurs. Il est aisé de démontrer qu'on a toujours

$$\max_i \min_j a_{ij} \leqslant \min_j \max_i a_{ij}.$$

Cela veut dire que la limite supérieure du gain est toujours supérieure au gain garanti. En effet, considérons la matrice-colonne p composée des éléments $p_i = \min_j a_{ij}$. Chacun des éléments de cette colonne est inférieur à l'élément homologue de toute autre colonne :

$$p_i \leqslant a_{ij}, \quad j = 1, \ldots, n.$$

Si le k-ième élément de p est le plus grand, on a tous les $j = 1, \ldots, n$

$$\max_i p_i = p_k \leqslant a_{kj} \leqslant \max_i a_{ij},$$

Etant donné que p_k est inférieur à chacun des n nombres, il est inférieur au plus petit d'entre eux :

$$\max_i p_i \leqslant \min_j \max_i a_{ij}.$$

ce qui coïncide avec l'inégalité qu'il fallait démontrer.

Les jeux pour lesquels se vérifie l'égalité

$$\max_i \min_j a_{ij} = \min_j \max_i a_{ij}$$

sont appelés *jeux à valeurs complètement déterminées*. Si $a_{i_0j_0}$ est l'élément égal aux deux membres de cette égalité, en choisissant la stratégie i_0 le premier joueur obtient son gain garanti qui est en même temps son gain maximal en cas de comportement prudent du second joueur. C'est le meilleur résultat auquel il peut espérer. De la même façon, le second joueur choisit la stratégie j_0 et gagne $-a_{i_0j_0}$. C'est la limite supérieure du gain du second joueur si le premier ne commet pas d'erreur. Si aucun des joueurs ne compte sur les erreurs de l'autre, le résultat du jeu est déterminé, d'où la dénomination de cette classe de jeux.

Les stratégies i_0 et j_0 sont appelées *stratégies optimales pures*.

Considérons à titre d'exemple un jeu à matrice

$$\begin{Vmatrix} 2 & 1 \\ 3 & 4 \end{Vmatrix}.$$

Il est clair que le premier joueur doit choisir la deuxième ligne, et le second, s'il ne compte pas sur l'erreur du premier, doit choisir la première colonne.

La situation change si l'on passe au jeu à matrice

$$\begin{Vmatrix} 3 & 1 \\ 2 & 4 \end{Vmatrix}.$$

Ce jeu n'est pas à valeurs déterminées vu que $\max_i \min_j a_{ij} = 2$ et $\min_j \max_i a_{ij} = 3$. En recherchant son gain garanti égal à 2, le premier joueur doit choisir la deuxième ligne. Si en même temps le second joueur recherche également son gain garanti égal à -3 et choisit la première colonne, il obtient le gain supérieur, égal à -2. On voit qu'avec ces stratégies le premier joueur se trouve dans une situation inférieure : il aurait pu obtenir 3 si en se fiant à la prudence du second il avait choisi la première ligne, auquel cas le second n'aurait obtenu que le minimum garanti. On voit que dans ce jeu il n'y a pas de stratégies optimales pures assurant les meilleurs résultats possibles aux deux joueurs qui jouent correctement.

6. Stratégies mixtes. Considérons un jeu pour lequel

$$a_{i_0j_0} = \max_i \min_j a_{ij} < \min_j \max_i a_{ij} = a_{i_1j_1}.$$

Jusque-là on a posé que le jeu n'intervenait qu'une fois. Dans cette hypo-

thèse on ne peut continuer l'examen du cas étudié. Supposons donc que le jeu se répète plusieurs fois.

Si dans ces multiples répétitions du jeu le premier joueur choisit constamment la stratégie i_0, le second joueur, en y comptant, choisira la stratégie j_0, de sorte que le premier obtiendra $a_{i_0j_0}$. Mais si le premier choisit contre toute attente la stratégie i_1, son gain sera alors $a_{i_1j_0}$, avec

$$a_{i_1j_0} \geqslant a_{i_1j_1} > a_{i_0j_0}.$$

Mais s'il continue à choisir la stratégie i_1, le second joueur, en le remarquant, lui laissera le gain min a_{i_1j} qui est inférieur et même strictement inférieur à $a_{i_0j_0}$.

Dans un jeu à valeurs non complètement déterminées la bonne tactique est de choisir une stratégie au hasard, de manière que l'adversaire ne puisse deviner la stratégie qui sera choisie la prochaine fois. Avec un choix correct des stratégies et de leurs probabilités, une telle décision permet d'élever le gain moyen au-dessus de $\max_i \min_j a_{ij}$ sans le faire toutefois monter jusqu'à $\min_j \max_i a_{ij}$.

Supposons que x^i, $i = 1, \ldots, m$, est la probabilité de choisir la stratégie i par le premier joueur et que y^j, $j = 1, \ldots, n$, est la probabilité de choisir la stratégie j par le second joueur. Il est naturel que

$$x^i \geqslant 0, \quad x^1 + \ldots + x^m = 1$$

et

$$y^j \geqslant 0, \quad y^1 + \ldots + y^n = 1.$$

Le gain du premier joueur représente maintenant une variable aléatoire dont l'espérance mathématique est facile à calculer. Vu que les stratégies sont choisies indépendamment, la probabilité de choisir un couple de stratégies (i, j) est x^iy^j et par suite, l'espérance mathématique du gain du premier joueur est

$$f(x, y) = \sum_{i,j} a_{ij}x^iy^j,$$

ou sous forme matricielle ${}^t\mathbf{x}A\mathbf{y}$.

Les matrices-colonnes $\mathbf{x} = {}^t\|x^1, \ldots, x^m\|$ et $\mathbf{y} = {}^t\|y^1, \ldots, y^n\|$ seront appelées *stratégies mixtes* du premier et du second joueur, tandis que les stratégies étudiées plus haut seront dites *pures* pour ne pas être confondues avec les stratégies mixtes. Les stratégies pures peuvent être considérées comme des stratégies mixtes définies par les colonnes de la matrice unité.

Du point de vue géométrique on peut se représenter les stratégies pures comme des vecteurs de base dans un espace m-dimensionnel (resp. n-dimensionnel), et les stratégies mixtes comme des combinaisons convexes de stratégies pures. Par exemple, l'ensemble de toutes les stratégies mixtes du premier joueur est représenté par l'enveloppe convexe de m points de l'espace affine m-dimensionnel, ces points n'appartenant à aucun plan $(m - 1)$-dimensionnel. Cet ensemble est le plus simple polyèdre de dimension m, appelé *simplexe* m-dimensionnel.

Etudions à quoi peuvent aboutir les joueurs s'ils utilisent les stratégies mixtes. Si le premier joueur choisit la stratégie x, il peut être certain que son gain sera supérieur à

$$\min_y {}^t x A y.$$

Ce minimum existe car la fonction ${}^t x A y$ est continue sur un ensemble fermé borné de valeurs de y pour un x fixé. Aussi le gain garanti du premier joueur ne dépasse-t-il pas pour les stratégies mixtes

$$\sup_x \min_y {}^t x A y.$$

De façon analogue, en choisissant la stratégie y, le second joueur gagne au moins $-\max_x {}^t x A y$ et, par suite, s'il joue correctement, le gain du premier joueur ne dépassera pas

$$\inf_y \max_x {}^t x A y.$$

On verra plus loin que ces bornes supérieure et inférieure peuvent être atteintes et, par suite, constituent un maximum et un minimum. Pour le moment, notons seulement qu'est vérifiée la suite d'inégalités suivante :

$$\min_y {}^t x A y \leqslant \sup_x \min_y {}^t x A y \leqslant \inf_y \max_x {}^t x A y \leqslant \max_x {}^t x A y. \qquad (8)$$

La démonstration n'est nécessaire que pour la deuxième inégalité. Il est évident que pour tous x et y est vérifiée la double inégalité

$$\min_y {}^t x A y \leqslant {}^t x A y \leqslant \max_x {}^t x A y.$$

Vu que le troisième membre de cette inégalité est indépendant de x on a

$$\sup_x \min_y {}^t x A y \leqslant \max_x {}^t x A y.$$

Dans cette inégalité le premier membre est indépendant de y. Donc,

$$\sup_x \min_y {}^t x A y \leqslant \inf_y \max_x {}^t x A y.$$

En fait, on a obligatoirement ici une égalité, ce qui constitue le théorème principal de la théorie des jeux matriciels qu'on démontrera plus loin. En attendant, l'inégalité obtenue est suffisante pour la démonstration de la formule (8).

Démontrons encore que pour tout x on a

$$\min_j \sum_i a_{ij} x^i \leqslant \min_y f(x, y).$$

En effet, la définition de $f(x, y)$ donne

$$f(x, y) = \sum_j y^j \sum_i a_{ij} x^i.$$

Si l'on substitue à chaque somme intérieure la plus petite d'entre elles, on obtient l'inégalité

$$f(x, y) \geqslant \left(\sum_j y^j\right) \min_j \sum_i a_{ij} x^i = \min_j \sum_i a_{ij} x^i.$$

Elle est vérifiée pour tous les y, en particulier pour celui qui réalise $\min_y f(x, y)$. L'inégalité est ainsi démontrée. On démontre de façon analogue que

$$\max_i \sum_j a_{ij} y^j \geqslant \max_x \sum_j a_{ij} x^i y^j.$$

7. Application de la programmation linéaire. Il découle des inégalités démontrées qu'en choisissant la stratégie x le premier joueur peut être certain que son gain sera supérieur à $v = \min_j \sum_i a_{ij} x^i$. La stratégie x satisfait au système d'inéquations

$$\begin{gathered} \sum_{i=1}^{m} a_{ij} x^i \geqslant v, \quad j = 1, \ldots, n, \\ x^1 + \ldots + x^m = 1 \\ x^1 \geqslant 0, \ldots, x^m \geqslant 0. \end{gathered} \tag{9}$$

Ceci étant, le premier joueur s'efforce de rechercher x de manière que v soit maximal. Pour réduire ce problème à un problème de programmation linéaire il faut transformer la matrice A. Plus précisément, ajoutons à tous les éléments de A un même nombre suffisamment grand pour que tous les éléments deviennent strictement positifs. Ceci est équivalent à l'attribution au premier joueur d'une somme pour la participation au jeu. Cette trans-

formation n'influe donc pas sur le choix de la stratégie par chacun des joueurs. Posons ainsi $\tilde{a}_{ij} = a_{ij} + \alpha$ et $\tilde{A} = \|\tilde{a}_{ij}\|$, où $\alpha > \max_{i,j} |a_{ij}|$. La somme v composée pour la matrice $\tilde{A}$ est strictement positive et on peut donc introduire des nouvelles variables

$$\xi^i = v^{-1} x^i, \quad i = 1, \ldots, m.$$

Le système d'inéquations (9) de la matrice A devient alors

$$\sum_{i=1}^{m} a_{ij} \xi^i \geqslant 1, \quad j = 1, \ldots, n,$$
$$\xi^1 + \ldots + \xi^m = v^{-1},$$
$$\xi^1 \geqslant 0, \ldots, \xi^m \geqslant 0.$$

Maintenant le désir de maximiser v conduit au problème suivant de programmation linéaire

$$\sum_{i=1}^{m} \tilde{a}_{ij} \xi^i \geqslant 1, \quad j = 1, \ldots, n,$$
$$\xi^i \geqslant 0, \quad i = 1, \ldots, m,$$
$$\xi^1 + \ldots + \xi^m \to \min.$$

Ce problème peut être écrit sous la forme matricielle si l'on introduit la matrice-colonne $\boldsymbol{i}$ dont tous les éléments sont égaux à 1. On ne précisera pas le nombre d'éléments de $\boldsymbol{i}$ qui peut être égal soit à m, soit à n selon le cas. Le problème pour le premier joueur prend la forme

$$\begin{gathered} {}^t\tilde{A}\boldsymbol{\xi} \geqslant \boldsymbol{i}, \quad \boldsymbol{\xi} \geqslant \boldsymbol{0}, \\ {}^t\boldsymbol{i}\boldsymbol{\xi} \to \min. \end{gathered} \tag{10}$$

Le problème est résoluble. En effet, la fonction économique est minorée car on a évidemment $v \leqslant \max_{x,y} {}^t x \tilde{A} y$. Le système des contraintes du problème (10) est compatible, car tous les $\tilde{a}_{ij} > 0$, et en choisissant $\xi^1 > \max_j \tilde{a}_{1j}^{-1}$, on rendra dans chaque inéquation un des termes plus grand que l'unité.

Construisons maintenant le problème pour le second joueur. Si le second joueur choisit la stratégie y, sa perte sera inférieure à $w = \max_i \sum_j \tilde{a}_{ij} y^j$. Ceci étant, on a

Ceci étant, on a

$$\sum_{j=1}^{n} \tilde{a}_{ij} y^j \leqslant w, \quad i = 1, \ldots, m,$$

$$y^1 + \ldots + y^n \geqslant 1,$$
$$y^1 \geqslant 0, \ldots, y^n \geqslant 0.$$

Les transformations analogues à celles qu'on a utilisées dans le problème pour le premier joueur réduisent le problème de minimisation de la perte maximale à la forme

$$\begin{gathered} {}'\boldsymbol{\eta}\,{}'\tilde{A} \leqslant {}'\boldsymbol{i}, \quad \boldsymbol{\eta} \geqslant 0, \\ {}'\boldsymbol{\eta}\,\boldsymbol{i} \to \max, \end{gathered} \tag{11}$$

où $\boldsymbol{\eta} = w^{-1}y$.

Il n'est pas difficile de remarquer que les problèmes du premier et du second joueur constituent un couple de problèmes duals. Le théorème 3 du § 3 et la formule (9) entraînent que les solutions $\boldsymbol{\eta}_0$ et $\boldsymbol{\eta}_0$ des problèmes (10) et (11) vérifient les égalités

$${}'\boldsymbol{\eta}_0\boldsymbol{i} = {}'\boldsymbol{\eta}_0({}'\tilde{A})\boldsymbol{\xi}_0 = {}'\boldsymbol{i}\,\boldsymbol{\xi}_0. \tag{12}$$

Passons aux anciennes variables x et y en tenant compte de ce que $\boldsymbol{\xi}_0$ et $\boldsymbol{\eta}_0$ correspondent à

$$v_0 = \min_j \sum_i \tilde{a}_{ij} x_0^i$$

et

$$w_0 = \max_i \sum_j \tilde{a}_{ij} y_0^j.$$

En multipliant (12) membre à membre par $v_0 w_0$, on obtient

$$\min_j \sum_i \tilde{a}_{ij} x_0^i = {}'y_0({}'\tilde{A})x_0 = \max_i \sum_j \tilde{a}_{ij} y_0^j.$$

On peut maintenant revenir à la matrice initiale A. Il vient

$$\min_j \sum_i (a_{ij} + \alpha) x_0^i = \min_j \sum_i a_{ij} x_0^i + \alpha \sum_i x_0^i = \min_j \sum_i a_{ij} x_0^i + \alpha$$

et

$$\max_i \sum_j (a_{ij} + \alpha) y_0^j = \max_i \sum_j a_{ij} y_0^j + \alpha,$$

ainsi que

$$\sum_{i,j} (a_{ij} + \alpha) x_0^i y_0^j = \sum_{i,j} a_{ij} x_0^i y_0^j + \alpha \sum_{i,j} x_0^i y_0^j = \sum_{i,j} a_{ij} x_0^i y_0^j + \alpha.$$

Donc,

$$\min_j \sum_i a_{ij} x_0^i = {}^t y_0 ({}^t A) x_0 = \max_i \sum_j a_{ij} y_0^j .$$

Il en découle immédiatement

$$\min_y {}^t y \, {}^t A x_0 = {}^t y_0 ({}^t A) x_0 = \max_x {}^t y_0 ({}^t A) x. \tag{13}$$

En effet, dans la double inégalité

$$\min_j \sum_i a_{ij} x_0^i \leqslant \min_y {}^t y \, {}^t A x_0 \leqslant {}^t y_0 ({}^t A) x_0$$

le premier membre est égal au troisième. Il en découle la première des égalités nécessaires, la seconde se démontre de façon analogue.

Maintenant les inégalités (8) et les égalités (13) entraînent de façon identique

$$\sup_x \min_y {}^t x A y = \inf_y \max_x {}^t x A y = {}^t x_0 A y_0 .$$

Il en découle non seulement l'égalité des bornes supérieure et inférieure mais aussi le fait que ces bornes sont accessibles. Le résultat obtenu peut être formulé sous forme de

THÉORÈME 2. *Pour tout jeu matriciel il existe des stratégies mixtes x_0 et y_0 pour lesquelles*

$$\max_x \min_y {}^t x A y = \min_y \max_x {}^t x A y = {}^t x_0 A y_0 .$$

Le nombre ${}^t x_0 A y_0$ est appelé *valeur du jeu*, et les stratégies x_0 et y_0 *stratégies optimales*. Le couple de stratégies optimales (x_0, y_0) porte le nom de *solution de jeu*.

Les stratégies optimales pures qui existent pour un jeu à valeurs complètement déterminées sont optimales en ce sens. On peut dire que le jeu est complètement déterminé si la valeur du jeu est égale à l'un des éléments de la matrice.

Les égalités (13) montrent qu'après avoir choisi sa stratégie optimale x_0 le premier joueur s'assure un gain égal à la valeur du jeu. Il ne peut obtenir davantage si le second joueur choisit aussi sa stratégie optimale. Si l'un des joueurs choisit sa stratégie optimale, le second, en s'abstenant de le faire, ne peut que rendre moins bon son gain.

Les stratégies optimales ne sont pas uniques, mais tout ce qui a été dit plus haut se rapporte à tout couple de stratégies optimales car chacune d'elles s'obtient d'une solution du problème correspondant de programma-

tion linéaire, et l'égalité (12) qui entraîne (13) a lieu pour tout couple de solutions des problèmes duals.

L'ensemble des stratégies optimales du premier joueur par exemple, est une partie convexe dans l'ensemble de ses stratégies mixtes, et chacune des stratégies optimales est une combinaison convexe de stratégies pures. Deux stratégies optimales peuvent contenir différentes stratégies pures. Il se peut d'ailleurs qu'une stratégie mixte se décompose de deux façons différentes en une combinaison convexe de stratégies pures.

Il ne faut pas oublier que la valeur du jeu est l'espérance mathématique du gain du premier joueur et que dans tous les cas on ne peut réaliser qu'un nombre fini de répétitions d'un jeu. Supposons que les joueurs ont adopté leurs stratégies optimales $\| x_0^1, \ldots, x_0^m \|$ et $\| y_0^1, \ldots, y_0^n \|$ et qu'à chaque répétition du jeu les stratégies pures i et j sont choisies avec des probabilités x_0^i et y_0^j respectivement. Alors le gain moyen du premier joueur tend vers la valeur du jeu lorsque le nombre de répétitions augmente. Tous les raisonnements sur les propriétés des stratégies optimales ne doivent être compris que dans ce sens-là.

Peut-on profiter des stratégies mixtes si le nombre de répétitions est petit ou même se limite à un seul jeu ? La réponse à cette question ainsi que l'étude détaillée de la théorie des jeux matriciels peuvent être trouvées dans le livre de Luce et Raiffe [25].

En conclusion on aurait voulu faire une remarque sur le lien logique de quelques théorèmes de ce chapitre. Un des principaux théorèmes est le théorème de Farkas sur les inéquations déduites du système d'inéquations linéaires. Le théorème de dualité en programmation linéaire s'ensuit en fait immédiatement vu que sa démonstration n'utilise, outre le théorème de Farkas, que le principe de Dedekind de continuité de l'ensemble des nombres réels. A leur tour, le théorème du flot maximal et le théorème principal de la théorie des jeux matriciels ont été obtenus par application directe du théorème de dualité. La démonstration de la proposition 15 du § 1, remplissant le rôle du lemme dans la démonstration du théorème de Farkas, est assez laborieuse. Mais, comme on le voit, les efforts dépensés à cette dernière permettent d'aboutir à des résultats remarquables.

BIBLIOGRAPHIE

1. Albert A. *Regression and the Moor-Penrose pseudoinverse*. Academic Press, N-Y, London, 1972.
2. Bellman R. *Introduction to matrix analysis*. McGron Hill, N-Y, Toronto, London, 1960.
3. Boulavski V., Zviaguina R., Yakovleva M. *Méthodes numériques de programmation linéaire*. Moscou, éd. Naouka, 1977 (en russe).
4. Chilov G. *Analyse mathématique. Espaces vectoriels de dimension finie*. Moscou, éd. Naouka, 1969 (en russe).
5. Danzig G.B. *Linear programming and extensions*. Princeton Univ. Press, Princeton N-Y, 1963.
6. Dreyfus M., Gangloff C. *La pratique du Fortran*. Dunod, Paris, 1975.
7. Efimov N. *Géométrie supérieure*. Moscou, éd. Mir, 1981 (traduit du russe).
8. Faddeev D.K., Faddeeva V.N. *Computational methods of linear algebra*. W.H. Freeman, San Francisco-London, 1963.
9. Fédoriouk M. *Equations différentielles ordinaires*. Moscou, éd. Naouka, 1980 (en russe).
10. Forsythe G.E., Moler C.B. *Computer solutions of linear algebraic systems*. Prentice-Hall, Eigenwood-Cliffs, N-Y, 1967.
11. Forsythe G.E., Malcolm M.A., Moler C.B. *Computer methods for mathematical computations*. Prentice-Hall, Eigenwood-Cliffs, N-Y, 1977.
12. Gantmacher F.R. *The theory of matrices*. Vol. I. II. Chelsea Publ. N-Y, 1959.
13. Glazman I., Lioubitch Iou. *Analyse linéaire finidimensionnelle*. Moscou, éd. Naouka, 1969 (en russe).
14. Golchtein E., Ioudine D. *Problèmes de transport en programmation linéaire*. Moscou, éd. Naouka, 1969 (en russe).
15. Hu T.C. *Interer programming and network flows*. Addison-Wesley, Reading, 1969.
16. Ikramov Kh. *Recueil de problèmes d'algèbre linéaire*. Moscou, éd. Mir, 1978 (traduit du russe).
17. Iline V., Pozniak E. *Géométrie analytique*. Moscou, éd. Mir, 1985 (traduit du russe).
18. Karmanov V. *Programmation mathématique*. Moscou, éd. Mir, 1977 (traduit du russe).
19. Kostrikin A. *Introduction à l'algèbre*. Moscou, éd. Mir, 1981 (traduit du russe).
20. Kostrikin A., Manine You. *Algèbre linéaire et géométrie*. Moscou, éd. MGU, 1980 (en russe).
21. Koudriavtsev L. *Analyse mathématique*. T. I, II. Moscou, éd. Vyschaïa chkola, 1981 (en russe).
22. Lankaster P. *Theory of matrices*. Academic Press, N-Y, London, 1969.
23. Lappo-Danilevski I. *Application des fonctions de matrices à la théorie des systèmes linéaires d'équations différentielles ordinaires*. Moscou, éd. Gostechizdat, 1957 (en russe).
24. *Linear inequalities and related topics* (Kuhn H.W., Tucker A.W. eds). Annals of math. study, N 38, Princeton Univ. Press, Princeton, N-Y, 1956.
25. Luce R., Raiffe H. *Games and decisions*. J. Willey, N-Y, 1957.
26. Maltsev A. *Eléments d'algèbre linéaire*. Moscou, éd. Naouka, 1970 (en russe).

27. Marcus M., Minc H. *A survey of matrix theory and matrix inequalities*. Alleyn and Bacon, Boston, 1964.
28. Motzkin T.S., Raiffe H., Thompson G.H., Throle R.M. *The double-description method*. In : *Contributions to the theory of games*, vol. II. Annals of math study, N 28, Princeton Univ. Press, Princeton, N-Y, 1953.
29. Nicaido H. *Convex structures and economic theory*. Acad. Press, N-Y, London, 1968.
30. Ostrowski A.M. *Solution of equations and systems of equations*. Acad. Press, N-Y, London, 1966.
31. Parodi M. *La localisation des valeurs caractéristiques des matrices et les applications*. Gauthier-Willars, Paris, 1959.
32. Postnikov M. *Leçons de géométrie 2e semestre. Algèbre linéaire et géométrie différentielle*. Moscou, éd. Mir, 1981 (traduit du russe).
33. Rozanov Iou. *Cours de théorie des probabilités*. Moscou, éd. Naouka, 1968 (en russe).
34. Samarski A. *Introduction à la théorie des schémas aux différences*. Moscou, éd. Naouka, 1971 (en russe).
35. Seber G.A.F. *Linear regression analysis*. J. Willey, N-Y, London, 1977.
36. Tchernikov S. *Inéquations linéaires*. Moscou, éd. Naouka, 1968 (en russe).
37. Twearson R.P. *Sparse matrices*. Academic Press, N-Y, London, 1973.
38. Tykhonov A., Arsénine V. *Méthodes de résolution de problèmes mal posés*. Moscou, éd. Mir, 1976 (traduit du russe).
39. Voïévodine V. *Algèbre linéaire*. Moscou, éd. Mir, 1976 (traduit du russe).
40. Voïévodine V. *Eléments de calcul numérique en algèbre linéaire*. Moscou, éd. Mir., 1980 (traduit du russe).
41. Voïévodine V. *Méthodes numériques en algèbre*. Moscou, éd. Naouka, 1966 (en russe).
42. Wilkinson J.H. *The algebraic eigenvalue problem*. Clanderon Press, Oxford, 1965.
43. Wilkinson J.H.,Reinsch G. *Handbuok for automatic computations. Linear algebra*. Springer-Verlag, Berlin, Heidelberg, N-Y, 1971.

INDEX DES NOMS

INDEX DES MATIÈRES

